On the Choice of Kernel and Labelled Data in Semi-supervised Learning Methods

Konstantin Avrachenkov[1], Paulo Gonçalves[2] and Marina Sokol[1]

[1] Inria Sophia Antipolis, 2004 Route des Lucioles, Sophia-Antipolis, France
[2] Inria Rhone-Alpes and ENS Lyon, 46 Allée Italie, Lyon, France

Abstract. Semi-supervised learning methods constitute a category of machine learning methods which use labelled points together with unlabelled data to tune the classifier. The main idea of the semi-supervised methods is based on an assumption that the classification function should change smoothly over a similarity graph, which represents relations among data points. This idea can be expressed using kernels on graphs such as graph Laplacian. Different semi-supervised learning methods have different kernels which reflect how the underlying similarity graph influences the classification results. In the present work, we analyse a general family of semi-supervised methods, provide insights about the differences among the methods and give recommendations for the choice of the kernel parameters and labelled points. In particular, it appears that it is preferable to choose a kernel based on the properties of the labelled points. We illustrate our general theoretical conclusions with an analytically tractable characteristic example, clustered preferential attachment model and classification of content in P2P networks.

1 Introduction

The first principal idea of the semi-supervised learning methods is to use few labelled points (points with known classification) together with the unlabelled data to tune the classifier. This drastically reduces the size of the training set. The second principal idea of the semi-supervised learning methods is to use a (weighted) similarity graph. If two data points are connected by an edge, this indicates some similarity of these points. Then, the weight of the edge, if present, reflects the degree of similarity. Later in the paper we show how the similarity graph can be constructed in a specific application. Each class has a classification function defined over all data points which gives a degree of relevance to the class for each data point. The third principal idea of the semi-supervised learning methods is that the classification function should change smoothly over the similarity graph. Intuitively, nodes of the similarity graph that are closer together in some sense are more likely to have the same label. This idea of classification function smoothness can be expressed using graph Laplacian or its modification. In particular, the authors of [14] proposed transductive learning, a semi-supervised learning method based on the Standard Laplacian. The authors of [13] and [15] used the Normalized Laplacian (or diffusion kernel). And the

authors of [3] used the Markov kernel. We observe that if one takes the method of [1] for detecting local cuts and takes seeds in [1] as the labelled data and considers sweeps as classification functions, then because the degrees of data points in different sweeps are the same, the resulting method will be equivalent to the semi-supervised method proposed in [3]. Recently in [5], the authors proposed a generalized optimization formulation which gives the above mentioned methods as particular cases. In the present work we provide more insights about the differences among the semi-supervised methods based on random walk theory, and give recommendations on how to choose the kernel and labelled points (of course, when there is some freedom in the choice of labelled points). It appears that the choice of labelled points influences the choice of kernel. In particular, we show that if the labelled points are chosen uniformly at random, the PageRank based method is the best choice for the semi-supervised kernel. On the other hand, if one can choose labelled points with large degrees or we know that labelled points given to us have large degrees, the Standard Laplacian method is the best choice.

The paper is organized as follows: In the next section we briefly describe the graph-based semi-supervised learning methods. We refer readers interested in more details on semi-supervised methods to several excellent surveys [8, 16, 17]. In Section 3 we provide general theoretical insights about semi-supervised learning methods and suggest how to choose the kernel and labelled points. Then, in Section 4 we illustrate the theoretical conclusions on an analytically tractable characteristic network example, on clustered preferential attachment model and with application to P2P content classification. In particular, for this specific application we show that with the right combination of labelled points and kernel one can achieve 95% precision with as little as 50 points per class for several hundred thousands unlabelled points. Finally, in Section 5 we give conclusions and provide directions for future research.

2 Semi-supervised learning methods

Suppose we need to classify N data points into K classes and assume P data points are labelled. That is, we know the class to which each labelled point belongs. Denote by V_k, the set of labelled points in class $k = 1, ..., K$. Thus, $|V_1| + ... + |V_K| = P$.

The graph-based semi-supervised learning approach uses a weighted graph connecting data points. The weight matrix, or similarity matrix, is denoted by W. Here we assume that W is symmetric and the underlying graph is connected. Each element $w_{i,j}$ represents the degree of similarity between data points i and j. Denote by D a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of matrix W: $d_i = \sum_{j=1}^{N} w_{i,j}$. Later in the paper we demonstrate how to construct the similarity matrix for a specific application.

Define an $N \times K$ matrix Y as

$$Y_{ik} = \begin{cases} 1, & \text{if } i \in V_k, \text{ i.e., point } i \text{ is labelled as a class } k \text{ point,} \\ 0, & \text{otherwise.} \end{cases}$$

We refer to each column Y_{*k} of matrix Y as a labeling function. Also define an $N \times K$ matrix F and call its columns F_{*k} classification functions. The 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 by means of the following optimization formulation [5]:

$$\min_F \{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|{d_i}^{\sigma-1}F_{i*} - {d_j}^{\sigma-1}F_{j*}\|^2 + \mu\sum_{i=1}^{N}{d_i}^{2\sigma-1}\|F_{i*} - Y_{i*}\|^2\}, \qquad (1)$$

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

The first order optimality condition gives explicit expressions for the classification functions

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

Once the classification functions are obtained, the points are classified according to the rule

$$F_{ik} > F_{ik'}, \forall k' \neq k \quad \Rightarrow \quad \text{Point } i \text{ is classified into class } k.$$

The ties can be broken in arbitrary fashion. We would like to note that our general scheme allows us to retrieve as particular cases:

- The Standard Laplacian method ($\sigma = 1$), [14]:

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

- The Normalized Laplacian method ($\sigma = 1/2$), [13]:

$$F_{*k} = \frac{\mu}{2+\mu}\left(I - \frac{2}{2+\mu}D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\right)^{-1} Y_{*k},$$

- The PageRank based method ($\sigma = 0$), [3]:

$$F_{*k} = \frac{\mu}{2+\mu}\left(I - \frac{2}{2+\mu}WD^{-1}\right)^{-1} Y_{*k}.$$

In the present work we try to answer the questions: which kernel (or which values of σ and μ) one needs to choose? and which points to label if we have some freedom with respect to labelling points? It turns out that these questions are not independent and one has to choose the kernel depending on the information available while labelling the points.

3 General theoretical considerations

First, let us transform the expression (2) to a more convenient form.

$$\begin{aligned} F_{*k} &= \frac{\mu}{2+\mu}\left(I - \frac{2}{2+\mu}D^{-\sigma}WD^{\sigma-1}\right)^{-1} Y_{*k} \\ &= \frac{\mu}{2+\mu}\left(D^{-\sigma}\left(I - \frac{2}{2+\mu}WD^{-1}\right)D^{\sigma}\right)^{-1} Y_{*k} \\ &= \frac{\mu}{2+\mu}D^{-\sigma}\left(I - \frac{2}{2+\mu}WD^{-1}\right)^{-1} D^{\sigma}Y_{*k}. \end{aligned}$$

Denoting $\alpha = 2/(2+\mu)$, transposing and using the fact that W is symmetric, we obtain

$$F_{*k}^T = (1-\alpha)Y_{*k}^T D^{\sigma}\left(I - \alpha D^{-1}W\right)^{-1} D^{-\sigma}. \tag{3}$$

Next we apply to the above expression the Blackwell series expansion [7, 12]

$$(1-\alpha)\left(I - \alpha D^{-1}W\right)^{-1} = \underline{1}\pi + (1-\alpha)H + \mathrm{o}(1-\alpha), \tag{4}$$

where π is the stationary distribution of the standard 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. We note that since 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 in the explicit form

$$\pi = (\underline{1}^T D\underline{1})^{-1}\underline{1}^T D. \tag{5}$$

Combining (3), (4) and (5), we can write

$$F_{*k}^T = (\underline{1}^T D\underline{1})^{-1}Y_{*k}^T D^{\sigma}\underline{1}\underline{1}^T D^{1-\sigma} + (1-\alpha)Y_{*k}^T D^{\sigma}HD^{-\sigma} + \mathrm{o}(1-\alpha).$$

In particular, we have

$$F_{ik} = \frac{d_i^{1-\sigma}}{\sum_{j=1}^N d_j}\sum_{p\in V_k} d_p^{\sigma} + (1-\alpha)d_i^{-\sigma}\sum_{p\in V_k} d_p^{\sigma}H_{pi} + \mathrm{o}(1-\alpha), \tag{6}$$

and, consequently, if $\sum_{p\in V_k} d_p^{\sigma} \neq \sum_{p\in V_{k'}} d_p^{\sigma}$ for some k and k', in the case when the parameter α is close to 1 (equivalently when μ is close to 0), then all points will be classified into the classes with the largest value of $\sum_{p\in V_k} d_p^{\sigma}$. An interesting exception is the case when $\sigma = 0$ and $|V_k| = const(k)$. In such a case, the zero order terms in the Blackwell expansions for the classification functions are the same for all classes and we need to compare the first order terms. Recall [10] that there is a connection between the mean first passage time of the standard random walk from node i to node j, m_{ij}, and the elements of the deviation matrix, namely, $m_{ij} = (\delta_{ij} + H_{jj} - H_{ij})/\pi_j$, where δ_{ij} is the Kronecker

delta. If $\sigma = 0$ and $|V_k| = const(k)$, substituting (6) into $F_{ik} - F_{ik'} > 0$ with $H_{pi} = H_{ii} - \pi_i m_{pi}$ for $i \neq p$ results in the condition

$$\sum_{s \in V_{k'}} m_{si} > \sum_{p \in V_k} m_{pi}.$$

This condition has a clear probabilistic interpretation: point i is classified into class k if the sum of mean passage times from the labelled points to point i is smallest for class k over all classes.

In addition to the standard random walk, it will also be helpful to consider a random walk with absorption $\{S_t \in \{1, ..., N\},\ t = 0, 1, ...\}$. At each step with probability α the random walk chooses next node among its neighbours uniformly and with probability $1 - \alpha$ goes into the absorbing state. The probabilities of visiting nodes before absorption given the random walk starts at node j, $S_0 = j$, are provided by the distribution

$$\mathrm{ppr}(j) = (1 - \alpha) e_j^T \left(I - \alpha D^{-1} W\right)^{-1}, \tag{7}$$

which is the personalized PageRank vector with respect to seed node j [9]. Here e_j denotes the j-th element of the standard basis.

Now we are ready to formulate the first result explaining the classification by the semi-supervised learning methods.

Theorem 1 *Data point i is classified by the generalized semi-supervised learning method* (1) *into class k, if*

$$\sum_{p \in V_k} d_p^\sigma q_{pi} > \sum_{s \in V_{k'}} d_s^\sigma q_{si}, \quad \forall k' \neq k, \tag{8}$$

where q_{pi} is the probability of reaching state i before absorption if $S_0 = p$.

Proof: Since $Y_{*k}^T = \sum_{p \in V_k} e_p^T$ and $F_{ik} = F_{*k}^T e_i$, from (3) we obtain

$$F_{ik} = \sum_{p \in V_k} d_p^\sigma (1 - \alpha) e_p^T \left(I - \alpha D^{-1} W\right)^{-1} e_i d_i^{-\sigma} = \frac{1}{d_i^\sigma} \sum_{p \in V_k} d_p^\sigma \mathrm{ppr}_i(p). \tag{9}$$

It has been shown in [6] that

$$\left(I - \alpha D^{-1} W\right)_{pi}^{-1} = q_{pi} \left(I - \alpha D^{-1} W\right)_{ii}^{-1},$$

where $(\cdot)_{pi}^{-1}$ denotes the (p, i)-element of the inverse matrix. Multiplying the above equation by $(1 - \alpha)$ yields

$$\mathrm{ppr}_i(p) = q_{pi} \mathrm{ppr}_i(i). \tag{10}$$

Thus, using relation (10) and equation (9), we conclude that for point i to be classified into class k we need

$$F_{ik} - F_{ik'} = \frac{\mathrm{ppr}_i(i)}{d_i^\sigma} \left(\sum_{p \in V_k} d_p^\sigma q_{pi} - \sum_{s \in V_{k'}} d_s^\sigma q_{si} \right) > 0, \quad \forall k' \neq k,$$

or, equivalently (8). □

Let us discuss the implications of Theorem 1. First, it is very interesting to observe that, using (8), one can decouple the effects from the choice of α and σ. A change in the value of α only influences the factor q_{pi} and a change in the value of σ only affects the factor d_p^σ. Second, the results of Theorem 1 are consistent with the conclusions obtained with the help of the Blackwell expansion. When α goes to one, q_{pi} goes to one and indeed classes with the largest value of $\sum_{p\in V_k} d_p^\sigma$ attract all points. Thus, the case of $\sigma = 0$ and $|V_k| = const(k)$ is especially interesting. In this case there is stability of classification even when α is close to one. Third, if $\sigma = 0$ and $|V_k| = const(k)$, one can expect that smaller classes will attract a larger number of "border points" than larger classes. Suppose that class k is smaller than class k'. Then, it is natural to expect that $q_{pi} > q_{si}$ with $p \in V_k$ and $s \in V_{k'}$. This observation will be confirmed by examples in the next section. This effect, if needed, can be compensated by increasing σ away from zero. And finally, fourth, we have the following rather surprising conclusion.

Corollary 1 *If labelled points have the same degree ($d_p = d$, $p \in V_k$, $k = 1, ..., K$), all considered semi-supervised learning methods provide the same classification.*

Now with the help of the following lemma, we can obtain another alternative condition for semi-supervised learning classification.

Lemma 1 *If the graph is undirected ($W^T = W$), then the following relation holds*

$$\mathrm{ppr}_j(i) = \frac{d_j}{d_i}\mathrm{ppr}_i(j). \tag{11}$$

Proof: We can rewrite (7) as follows

$$\mathrm{ppr}(i) = (1-\alpha)e_i^T[D - \alpha W]^{-1}D,$$

and hence,

$$\mathrm{ppr}(i)D^{-1} = (1-\alpha)e_i^T[D - \alpha W]^{-1}.$$

Since matrix W is symmetric, $[D - \alpha W]^{-1}$ is also symmetric and we have

$$[\mathrm{ppr}(i)D^{-1}]_j = (1-\alpha)e_i^T[D-\alpha W]^{-1}e_j = (1-\alpha)e_j^T[D-\alpha W]^{-1}e_i = [\mathrm{ppr}(j)D^{-1}]_i.$$

Thus, $\mathrm{ppr}_j(i)/d_j = \mathrm{ppr}_i(j)/d_i$, which completes the proof. □

Theorem 2 *Data point i is classified by the generalized semi-supervised learning method* (1) *into class k, if*

$$\sum_{p\in V_k} \frac{\mathrm{ppr}_p(i)}{d_p^{1-\sigma}} > \sum_{s\in V_{k'}} \frac{\mathrm{ppr}_s(i)}{d_s^{1-\sigma}}, \quad \forall k' \neq k. \tag{12}$$

Proof: Follows from equation (9) and Lemma 1. □

We note that in the statement of Theorem 2 the "reversed" PageRank is used instead of the PageRank in (9). In particular, this provides another interesting interpretation of the PageRank based method. If we set $\sigma = 0$ in (12), it appears that we need to compare the reversed PageRanks divided by the degrees of the labelled points. As already mentioned in the Introduction, if one considers the sweeps from [1] as classification functions, then the degrees of the nodes to be classified are cancelled in the sweeps. However, if we now view the PageRank method in terms of the reversed PageRank, the division by the degree of the PageRank values remains essential. This provides another interesting interpretation of sweeps defined in [1].

4 Evaluation

Let us illustrate the theoretical results with the help of a characteristic network example, clustered preferential attachment graph and application to P2P content classification.

Characteristic network example: Let us first consider an analytically tractable network example. Despite its simplicity, it clearly demonstrates major properties of graph-based semi-supervised learning methods. There are two classes, A and B with $|A| = N_1$ and $|B| = N_2$. Each class is represented by a star network. The two classes are connected by a link connecting two leaves. The graph of the model is given in Figure 2(a).

The central nodes with indices 1 and $N_1 + N_2$ are the obvious choice for labelled points. In order to determine the classification functions analytically, we need to calculate the matrix $Z = [I - \alpha D^{-1}W]^{-1}$. It is easier to calculate the symmetric matrix $C = [D - \alpha W]^{-1}$. Once the matrix C is calculated, we can immediately retrieve the elements of matrix Z by the formula

$$Z_{ij} = C_{ij} d_j. \tag{13}$$

Thus we need to solve a system of equations $[D - \alpha W]C_{*,j} = e_j$. Since we have chosen the central nodes as labelled points and due to the symmetry of the graph, we actually need to solve only one system for $j = 1$ of six equations

$$\begin{aligned}
(N_1 - 1)C_{1,1} - (N_1 - 2)\alpha C_{2,1} - \alpha C_{N_1,1} &= 1 \\
C_{2,1} &= \alpha C_{1,1} \\
C_{N_1-1,1} &= \alpha C_{1,1} \\
-\alpha C_{1,1} + 2C_{N_1,1} - \alpha C_{N_1+1,1} &= 0 \\
-\alpha C_{N_1,1} + 2C_{N_1+1,1} - \alpha C_{N_1+N_2,1} &= 0 \\
C_{N_1+2,1} &= \alpha C_{N_1+N_2,1} \\
-\alpha C_{N_1+1,1} - (N_2 - 2)\alpha C_{N_1+2,1} + (N_2 - 1)C_{N_1+N_2,1} &= 0
\end{aligned}$$

Solving the above system, in particular, we obtain

$$C_{N_1,1} = \frac{\alpha(2N_2 - 2 - \alpha^2(2N_2 - 3))}{R}, \tag{14}$$

$$C_{N_1+1,1} = \frac{\alpha^2(N_2 - 1 - \alpha^2(N_2 - 2))}{R}, \tag{15}$$

with

$$R = (1-\alpha^2)(-2\alpha^4 N_2 - 2\alpha^4 N_1 + 4\alpha^4 + \alpha^4 N_2 N_1 - 9\alpha^2$$
$$+7\alpha^2 N_2 + 7\alpha^2 N_1 - 5N_2\alpha^2 N_1 + 4N_2 N_1 + 4 - 4N_1 - 4N_2).$$

Consider first the PageRank based method ($\sigma = 0$). According to the theoretical consideration, it is very likely that some points will be misclassified into a smaller class. Suppose that $N_1 < N_2$ and consider border points. The point $N_1 + 1$ will be classified into class B by the PageRank based method if and only if

$$\frac{Z_{1,N_1+1}}{Z_{N_1+N_2,N_1+1}} = \frac{C_{1,N_1+1}}{C_{N_1+N_2,N_1+1}} < 1.$$

Using slightly more convenient notation $n_i = N_i - 1, i = 1, 2$, we can rewrite the above condition as follows:

$$\frac{\alpha(n_2 - \alpha^2(n_2 - 1))}{2n_1 - \alpha^2(2n_1 - 1)} < 1,$$

or, equivalently, $(1 - n_2)\alpha^2 + (2n_1 - n_2)\alpha + 2n_1 > 0$. If $2n_1 + 1 > n_2$, the above inequality holds for any $\alpha \in (0, 1)$. And consequently, for any $\alpha \in (0, 1)$ the point $N_1 + 1$ is classified into class B. However, if $2n_1 + 1 < n_2$ (class A is significantly smaller than class B), for $\alpha \in (\bar{\alpha}, 1)$ point $N_1 + 1$ will be erroneously classified into class A. The expression for $\bar{\alpha}$ is given by

$$\bar{\alpha} = \frac{-(n_2 - 2n_1) + \sqrt{(2n_1 + n_2)^2 - 8n_1}}{2(n_2 - 1)}.$$

If we fix the value of n_1 and let n_2 go to infinity, we get $\bar{\alpha} \to 0$. Thus, if the sizes of A and B are very different, the point $N_1 + 1$ will be misclassified for nearly all values of the parameter α.

Now we analyse the performance of the Standard Laplacian method ($\sigma = 1$). According to the general theoretical considerations, the Standard Laplacian method has a tendency to classify more points into a larger class. We consider the classification of the point with index N_1 (still assuming $N_1 < N_2$). It will be classified correctly if and only if

$$\frac{Z_{N_1,1}}{Z_{N_1,N_1+N_2}} > 1,$$

or, equivalently,

$$\frac{n_1(2n_2 - \alpha^2(2n_2 - 1))}{n_2\alpha(n_1 - \alpha^2(n_1 - 1))} > 1$$

which results in the following cubic inequality

$$\alpha^3 n_2(n_1 - 1) - \alpha^2 n_1(2n_2 - 1) - \alpha n_2 n_1 + 2n_2 n_1 > 0.$$

Consider a linear scaling $n_2 = Kn_1, K > 1$. Then, the above inequality can be rewritten in the form

$$\alpha^3 \left(1 - \frac{1}{n_1}\right) - \alpha^2 \left(2 - \frac{1}{Kn_1}\right) - \alpha + 2 > 0.$$

This inequality can be regarded as a regularly perturbed inequality with respect to $1/n_1$ (see e.g., [2]). If we let n_1 go to infinity, the limiting inequality can be easily factored, i.e.,$(1-\alpha)(1+\alpha)(2-\alpha) > 0$. Since the perturbation is regular, when n_1 varies in the vicinity of infinity the roots change slightly. In particular, using the implicit function theorem, we can find that the root near 1 changes as follows:

$$\bar{\bar{\alpha}} = 1 - \frac{K-1}{2K}\frac{1}{n_1} + \mathrm{o}\left(\frac{1}{n_1}\right).$$

In particular, this means that if the sizes of classes are large, the Standard Laplacian method performs well for nearly all values of α from the interval $(0, 1)$. This is in contrast with the PageRank based method.

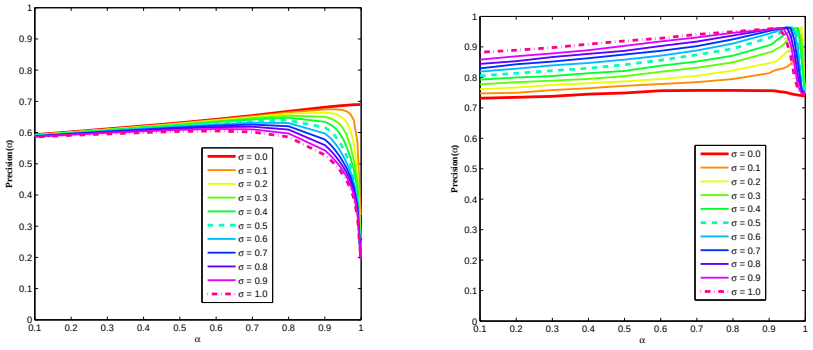
We summarize and illustrate various considered cases by means of numerical examples presented in Table 1. Our main conclusion from this characteristic network model is that the PageRank based method is a safe choice as it can misclassify at most one point in this particular example whereas with α close to one the Standard Laplacian method can classify all points in the largest class. On the other hand if parameter α is chosen appropriately, the Standard Laplacian method gives a perfect classification for nearly all values of α, even when classes have many points and very different sizes.

N_1	N_2	PR	SL
20	100	$v_{N_1+1} \mapsto A$ if $\alpha \geq \bar{\alpha} = 0.3849$	$v_{N_1} \mapsto B$ if $\alpha \geq \bar{\bar{\alpha}} = 0.9803$, $A \mapsto B$ if $\alpha \geq 0.9931$
20	200	$v_{N_1+1} \mapsto A$ if $\alpha \geq \bar{\alpha} = 0.1911$	$v_{N_1} \mapsto B$ if $\alpha \geq \bar{\bar{\alpha}} = 0.9780$, $A \mapsto B$ if $\alpha \geq 0.9923$
200	2000	$v_{N_1+1} \mapsto A$ if $\alpha \geq \bar{\alpha} = 0.1991$	$v_{N_1} \mapsto B$ if $\alpha \geq \bar{\bar{\alpha}} = 0.9978$, $A \mapsto B$ if $\alpha \geq 0.9992$

Table 1. Comparison between different methods in terms of classification errors

Clustered Preferential Attachment model: Let us now consider a synthetic graph generated according to the clustered preferential attachment model. Our model has 5 unbalanced classes (1500 / 240 / 120 / 100 / 50). Once a node is generated, it has two links which it attaches independently with probability 0.98 within its class and with probability 0.02 outside its class. In both cases a link is attached to a node with probability proportional to the number of existing links. First, we test the case of random labelled points. Five labelled points were chosen randomly for each class and results are averaged over 100 realizations. The precision of classification for various values of σ and α is given in Figure 1(a). Then, in each class we have chosen 5 labelled points with maximal degrees. The results of classification are given in Figure 1(b). We obtain conclusions consistent with the characteristic network model. If no information is available

for assignment of the labelled points, the PageRank method is a safe choice. If one can choose labelled points with large degrees, it is better to use the Standard Laplacian method. There could be a significant gain in precision (roughly from 70% to 95%). It can be observed that the Standard Laplacian method is not too sensitive to the value of α if we stay well away from $\alpha = 1$.

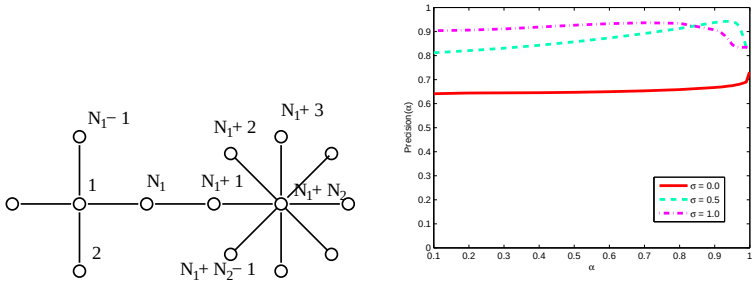


(a) Random Labelled Points (b) Max Degree Labelled Points

Fig. 1. Clustered Preferential Attachment Model: Precision of classification.

Application to P2P content classification: Finally, we would like to conclude the illustration with an application to P2P content classification. For lack of space, here we just very briefly outline the experiment and the results. An interested reader can find more details about this application in [4]. Using the technology developed in [11] we had an access to all world-wide Torrents managed by BitTorrents protocol. In particular, within one week we could observe 200413 different content files. Each file is a data point and we create an edge between two data points i and j if the same user downloaded two files i and j. By such a construction, graph has 50726946 edges. Consider an example of classification of the content by language (e.g., language of a movie or language of a book). Fortunately, a big portion of the content is tagged, so we can compare with the ground truth for some content. We have chosen to classify the content according to five major languages (English, French, Italian, Japanese, Spanish). For each language we have chosen 50 labelled points with the maximal degree within the ground truth points. Since we do not have ground truth for all the points, it is assumed that choosing random points from the ground truth will not be representative (popular content is more likely to be tagged). The precision of classification for $\sigma = 0.0; 0.5; 1.0$ and various values of α is given in Figure 2(b). The figure is consistent with Figure 1(b). In Tables 2 and 3 we provide cross-validation matrices for the Standard Laplacian and PageRank based methods with $\alpha = 0.8$. We can observe that as in the previous examples, the PageR-

ank method pulls elements from the largest class to the smaller classes and the Standard Laplacian method does the opposite. Thus, in the case of unbalanced classification, by choosing σ, one admits a trade off between precision and recall for smaller classes.



(a) Characteristic network model. (b) P2P Content Classification.

Fig. 2.

Classified as→	En	Fr	It	Jp	Sp
English	36097	22	134	53	159
French	903	909	7	1	4
Italian	308	1	2123	1	17
Japanese	583	7	4	120	6
Spanish	662	1	14	0	1804

Table 2. $\sigma = 1.0$, Precision 93.43%

Classified as→	En	Fr	It	Jp	Sp
English	22276	3812	3095	6233	1049
French	87	1618	38	63	18
Italian	24	27	2329	40	30
Japanese	45	43	25	568	39
Spanish	124	78	83	52	2144

Table 3. $\sigma = 0.0$, precision 65.85%

5 Conclusion, future research and acknowledgements

Using random walk theory, we provide insights about different graph-based semi-supervised learning methods. We also suggest the following recommendations. If possible, choose labelled points with large degrees. Then, adopt the Standard Laplacian method with α in the upper-middle range of the interval $(0, 1)$. If finding large degree points is not feasible or recall is more important than precision for small classes, choose the PageRank based method. In our near future research we plan to study in more detail the choice of the regularization parameter.

This research is funded by Inria Alcatel-Lucent Joint Lab. We also would like to thank P.G. Howlett, J.K. Sreedharan and anonymous reviewers whose comments helped to improve the presentation of the results.

References

1. R. Andersen, F. Chung, and K. Lang. Using pagerank to locally partition a graph. *Internet Mathematics*, 4(1):35–64, 2007.
2. K. Avrachenkov. *Analytic Perturbation Theory and its Applications, PhD Thesis*. University of South Australia, Adelaide, Australia, 1999.
3. K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S.K. Pham, and E. Smirnova. Pagerank based clustering of hypertext document collections. In *Proceedings of the 31st annual international ACM conference on research and development in information retrieval*, SIGIR '08, pages 873–874. ACM, 2008.
4. K. Avrachenkov, P. Gonçalves, A. Legout, and M. Sokol. Classification of content and users in bittorrent by semi-supervised learning methods. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2012 8th International, Workshop on Traffic Analysis and Classification*, pages 625–630, 2012.
5. K. Avrachenkov, P. Gonçalves, A. Mishenin, and M. Sokol. Generalized optimization framework for graph-based semi-supervised learning. In *Proceedings of SIAM Conference on Data Mining (SDM'2012), 9 pages*, 2012.
6. K. Avrachenkov and N. Litvak. The effect of new links on google pagerank. *Stochastic Models*, 22(2), 2006.
7. D. Blackwell. Discrete dynamic programming. *Ann. Math. Statist.*, 33:719–726, 1962.
8. Z. Guo, Z. Zhang, E.P. Xing, and C. Faloutsos. Semi-supervised learning based on semiparametric regularization. In *SDM'08 Proceedings*, pages 132–142, 2008.
9. T.H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*, pages 517–526, 2002.
10. J.G. Kemeny and J.L. Snell. *Finite Markov chains*. Springer, 1st edition, 1976.
11. S. Le Blond, A. Legout, F. Lefessant, W. Dabbous, and M.A. Kaafar. Spying the world from your laptop: identifying and profiling content providers and big downloaders in bittorrent. In *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, LEET'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.
12. M.L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.
13. D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, 2004.
14. D. Zhou and C.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.
15. D. Zhou and B. Schölkopf. A regularization framework for learning from graph data. In *Proceedings of the Workshop on Statistical Relational Learning at Twenty-first International Conference on Machine Learning, (ICML'2004), Canada, 6 pages*, 2004.
16. X. Zhu. Semi-supervised learning literature survey, technical report 1530, department of computer sciences, university of wisconsin, madison, 2005.
17. X. Zhu and A.B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.