# Supervised classification and visualization of social networks based on a probabilistic latent space model

**Charles Bouveyron** CHARLES.BOUVEYRON@UNIV-PARIS1.FR
SAMOS-MATISSE, CES, UMR CNRS 8174, Université Paris 1 (Panthéon-Sorbonne), Paris, FRANCE

**Hugh Chipman** HUGH.CHIPMAN@ACADIAU.CA
Department of Mathematics and Statistics, Acadia University, Wolfville, CANADA

**Etienne Côme** ETIENNE.COME@UNIV-PARIS1.FR
SAMOS-MATISSE, CES, UMR CNRS 8174, Université Paris 1 (Panthéon-Sorbonne), Paris, FRANCE

**Keywords**: social networks, supervised classification, visualization, latent space model.

## Abstract

Graph-structured networks are widely used to represent relationships between persons in organizations or communities. Recently, the need of classifying and visualizing such data has suddenly grown due to the emergence of a large number of social network websites. We propose in this paper two supervised approaches for learning a latent space model of the network taking into account both the observed class labels and the graph structure. The first proposed approach introduces the class information through the conditional model of the link existence between two nodes whereas the second one considers the class labels as new observed variables. The learned models are then used to project and classify new nodes.

## 1. Introduction

Increasingly, it is becoming possible to observe "network information" in a variety of contexts, such as email transactions, connectivity of webpages, protein-protein interactions and social networking. A number of scientific goals can apply to such networks, ranging from unsupervised problems such as describing network structure, to supervised problems such as predicting node labels with information on their relationships. In this paper we extend an unsupervised model proposed by Hoff *et al.* (Hoff et al., 2002), to deal with supervised classification problems. The "network structure" in our data are binary relations observed between every pair of nodes in the network. The response for the supervised problem will be categorical, although other response types are possible. Social network classification could have many applications, such as categorization of web pages into topics according to their link structure or classification of social networks into subcommunities for marketing purposes.

Section 2 presents the latent social network model on which our work is based. Section 3 presents the proposed supervised approaches. Experimental results on a real dataset are presented in Section 4 highlighting the main features of the proposed supervised latent models.

## 2. The latent space model

Among existing probabilistic social network models, we choose to start with the latent space model proposed by Hoff *et al.* in (Hoff et al., 2002). This model provides probabilistic inference for the visualization and analysis of a social network. A social network is usually represented by a $n \times n$ socio-matrix where its elements $Y_{ij}$ denotes the existing relation between the nodes $i$ and $j$, for $i, j = 1, ..., n$. As in (Hoff et al., 2002), we focus on binary-valued relations, *i.e.*, $Y_{ij} \in \{0, 1\}$ for $i, j = 1, ..., n$. Let $Y_{ij}$ take the value $1$ if a tie exists between the node $i$ and $j$ and $0$ otherwise. For example, later in Section 4 we consider data in which $Y_{ij} = 1$ indicates friendship between individuals $i$ and $j$. The latent space model assumes in addition that the presence or the absence of a tie between two nodes is independent of the other ties in the network, conditional on the the locations of the nodes in the latent space. The latent coordinates $Z_i$, $i = 1, ..., n$, are assumed to be $p$-dimensional where $p$ is unknown. The latent space model representing

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

socio-matrix $Y$ takes the following form:

$$logit(P(Y_{ij} = 1|\theta)) = \alpha - \|Z_i - Z_j\|,$$

where $logit(P) = \log(P/(1-P))$, $\theta = \{\alpha, Z\}$ are the parameters of the model, $\alpha$ determines the prior probability of an existing link between two nodes and $Z_i$ is the position of the $i$th node in the $p$-dimensional latent space. Thus, using this model, nodes $i$ and $j$ have a high probability to be connected if $\alpha$ is large or if they are close in the latent space, *i.e.*, $\|Z_i - Z_j\|$ is close to $0$. Given estimated values of parameters $\alpha, Z_1, \ldots, Z_n$ it is possible to predict the existence of a tie between two nodes using the distance between them in the latent space.

To learn the latent space model, we must estimate $\alpha$ and $Z_i, \ldots, Z_n$ for a fixed value of latent space dimension $p$. Dimension $p$ can be chosen by cross-validation or a criterion such as BIC. We use a Maximum Likelihood (ML) approach for the estimation of the model parameters, for computational ease. Alternatives such as Bayesian estimation are discussed in (Hoff et al., 2002). The log-likelihood can be expressed as follows:

$$\log(L(\theta)) = \sum_{i \neq j} \left[ y_{ij}\eta_{ij} - \log(1 + \exp(\eta_{ij})) \right],$$

where $\eta_{ij} = -\alpha + \|Z_i - Z_j\|$. We use simulated annealing to maximize the log likelihood and thus estimate $\theta = \{\alpha, Z\}$.

## 3. Supervised classification in a latent space

We propose in this section two supervised approaches for the classification of latent networks.

### 3.1. Supervised latent model SL1

As with any supervised classification approach, the proposed method consists in two phases: a learning phase and a classification phase.

**Learning phase** This phase aims to first learn a latent model which takes into account the class information and then learn a supervised classifier in the resulting latent space. The main idea of this appraoch is to introduce the supervised information within the latent space model through a covariate term $\beta X_{ij}$. This covariate term exists in the original latent space model (Hoff et al., 2002) but is usually not used ($\beta$ is set to $0$). The supervised latent model SL1 has the following form:

$$logit(P(Y_{ij} = 1|\theta)) = \alpha - \beta X_{ij} - \|Z_i - Z_j\|,$$

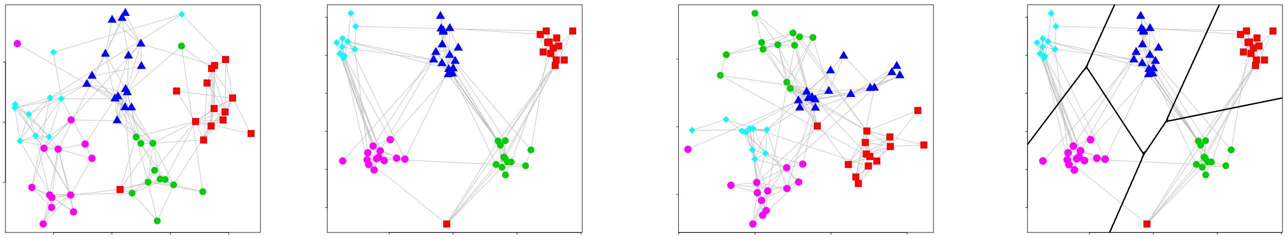
where $X_{ij}$ is equal to $1$ it the nodes $i$ and $j$ are in the same class and $-1$ if they are not. The parameter $\beta$ is an hyperparameter which tunes the importance given to the supervision in the model. Particularly, the model SL1 reduces to the classical latent model if $\beta$ is equal to $0$. Inclusion of $\beta X_{ij}$ forces model SL1 to provide latent positions which respect the class memberships. Indeed, two nodes from the same class will have to be close in the latent space whereas nodes from different classes will have to be far away. Once the latent model parameters are estimated, it is possible to learn a supervised classifier in the latent space associated with the LSN model. Since our approach is very general and does not make specific assumptions, the supervised classifier can be either generative or discriminative.

**Classification phase** Once the latent space and the classifier are learned, it is possible to classify new nodes using their observed links with learning nodes. The classification phase is also a two-step approach: the new node has to be projected in the learned latent space before being assigned to one of the classes. Indeed, before we can predict the supervised response associated with a new node, we need to know the value of its position $Z$ in the latent space. It is therefore necessary to project the new nodes in the latent space in which the classifier was learned. However, in the learning phase, we learned only the coordinates of the nodes in the latent space and not a basis of this latent space. It is thus not possible to directly project the new nodes in the learned latent space. Instead, we use observed links between the new nodes and the learning nodes to position the new nodes in the already-learned latent space. We estimate the latent positions of new nodes by maximizing the likelihood of the whole dataset, *i.e.*, the learning and the test datasets, as a function of only the latent positions $Z$ of the new nodes. In this step we estimate only the latent positions of the new nodes and the positions of the learning nodes remain fixed. As starting value of the new $Z$, we use the average latent position of all nodes connected with new node $i$. Our projection method can simultaneously find the latent position of several new nodes. Indeed, finding the latent positions of $m$ new nodes is equivalent to find the maximum of the likelihood surface in a $(mp+1)$-dimensional space. An additional challenge is that for the test set, the $X_{ij}$ are unknown, since they depend on unobserved labels. Setting $X_{ij} = 0$ for indices $i$ or $j$ in the test set will represent this missing information. After the projection of the test nodes, it is easy to assign the new nodes to one of the $k$ classes using the learned classifier according to their positions in the latent space.

### 3.2. Supervised latent model SL2

Another solution to deal with labeled data in the latent space model described previously is to define a conditional relationship between latent positions and classes.

**Learning phase** We propose to model this relationship using a logistic regression model. With such an approach, the conditional probability that sample $i$ comes from class



(a) Usual latent space (b) Supervised latent space (SL1) (c) Supervised latent space (SL2) (d) Sup. classifier (LDA) on SL1

*Figure 1.* Supervised classification and visualization for the Add-Health dataset.

$h$ given its latent position $Z_i$ is:

$$P(C_{ih}=1|z_i)=\frac{\exp(\beta_h^t\, Z_i)}{1+\sum_{l=1}^{k-1}\exp(\beta_l^t\, Z_i)}, \text{if } h<k$$

$$P(C_{ik}=1|z_i)=\frac{1}{1+\sum_{l=1}^{k-1}\exp(\beta_l^t\, Z_i)},$$

where $k$ is the number of classes and the $C_{ih}$ are dummy variables encoding class membership: $C_{ih}=1$ if node $i$ belongs to class $h$ and $C_{ih}=0$ otherwise. By using the additional assumption that the link variables $Y_{ij}$ described previously are independent from classes of node $i$ and $j$ when their latent positions are known, we can build the following log-likelihood criteria which takes into account both the known class membership of nodes and the graph structure:

$$\log(L(\theta))=\sum_{i\neq j}[y_{ij}\eta_{ij}-\log(1+\exp(\eta_{ij}))]$$
$$+\sum_{i=1}^{n}\sum_{h=1}^{k-1}c_{ih}\left[\beta_h^t\, Z_i-\log\left(1+\sum_{l=1}^{k-1}\exp(\beta_l^t\, Z_i)\right)\right],$$

The graph nodes can be therefore projected on the latent space by optimizing this log-likelihood with respect to the latent positions of nodes and the others parameters.

**Classification phase** As with the previous method, new nodes can be projected on the latent space by maximizing the likelihood of the whole dataset, *i.e.*, the learning and the test datasets, as a function of only the latent positions $Z$ of the new nodes. The same optimization method as in method SL1 can be used to estimate the final $Z$. When the new nodes position have been found, according to their relationship with base nodes, their classes can simply be estimated with the help of the logistic regression model already fitted in the learning phase.

## 4. Numerical experiments

The social network studied here is from the National Longitudinal Study of Adolescent Health (“Add-Health”). The data were collected in 1994-95, at 80 high schools and 52 middle schools in the USA. The whole study is detailed in (Harris, K. *et al.*, 2003). In addition to personal and social informations, each student was asked to nominate his best friends. We consider here the social network constructed with the answers of 71 students from a single school. Two adolescents who nominated nobody were removed from the network. For this experiment, we use the grade of each student as classes. Figure 1 presents the usual latent space (a), the two supervised latent spaces (b,c) and the learned classifier in supervised latent space SL1 (d). Both latent representations give a clear understanding of the relationships within this school: different classes are arranged from left to right according to their associated grades. However, we observe that supervised latent spaces provide a clearer visualization and are actually well suited for classifying new nodes. We do not present results on projection and classification of new nodes, but both are satisfying. This experiment illustrates that our approach enables the visualization and classification of nodes for a quite complex network (69 nodes and 5 classes).

## 5. Further work

The supervised classification methods introduced in this work require that the latent space dimension be selected. We think that this could be done using a Bayesian approach, such as BIC. The representation of a unconventional network as a more conventional $n \times p$ feature matrix also opens up additional possibilities, such as the use of additional covariates that summarize network activity. Finally, this approach could be extended to “dynamic network” situations, in which network structure is changing over time.

## References

Harris, K. *et al.* (2003). *The national longitudinal of adolescent health: research design* (Technical Report). Carolina Population Center, University of North Carolina.

Hoff, P., Raftery, A., & Handcock, M. (2002). Latent spaces approaches to social network analysis. *Journal of the American Statistical Association*, *97*, 1090–1098.