Semi-supervised Learning Methods for Classification and Community Detection

Konstantin Avrachenkov

Based on join works with
Marina Sokol (Inria), Alexey Mishenin (SPSU),
Paulo Gonçalves (Inria) and Arnaud Legout (Inria).

Nizhniy Novgorod, 18 May 2015

Semi-supervised vs Supervised Learning

In the supervised learning the data are divided into training set and unclassified set.

A classifier is first tuned on the training set and then it is applied for classification of the raw data.

Often, expert classification of a large training set is expensive and might even be infeasible.

Semi-supervised vs Supervised Learning

The main idea of the semi-supervised learning approach is to create a synergy between the labelled and unlabelled data.

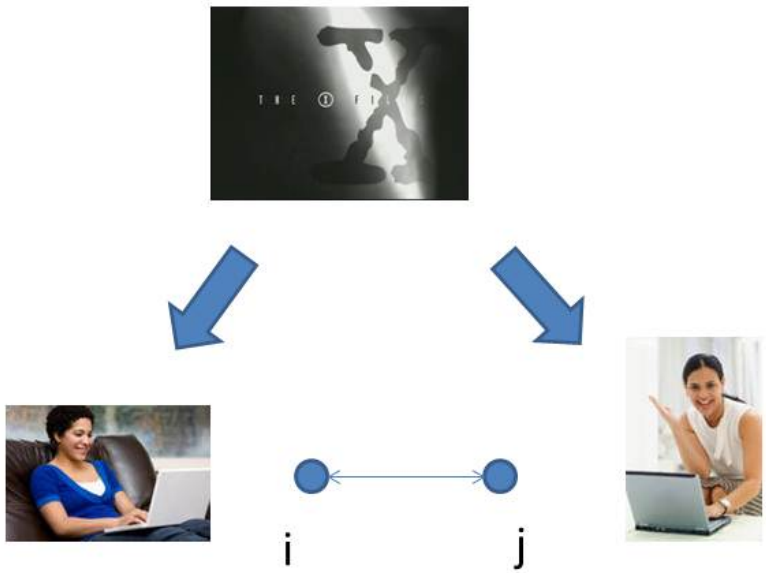
One large class of semi-supervised learning algorithms is based on the use of the similarity graph.

Similarity graph

Often the similarity graph is naturally provided by application.

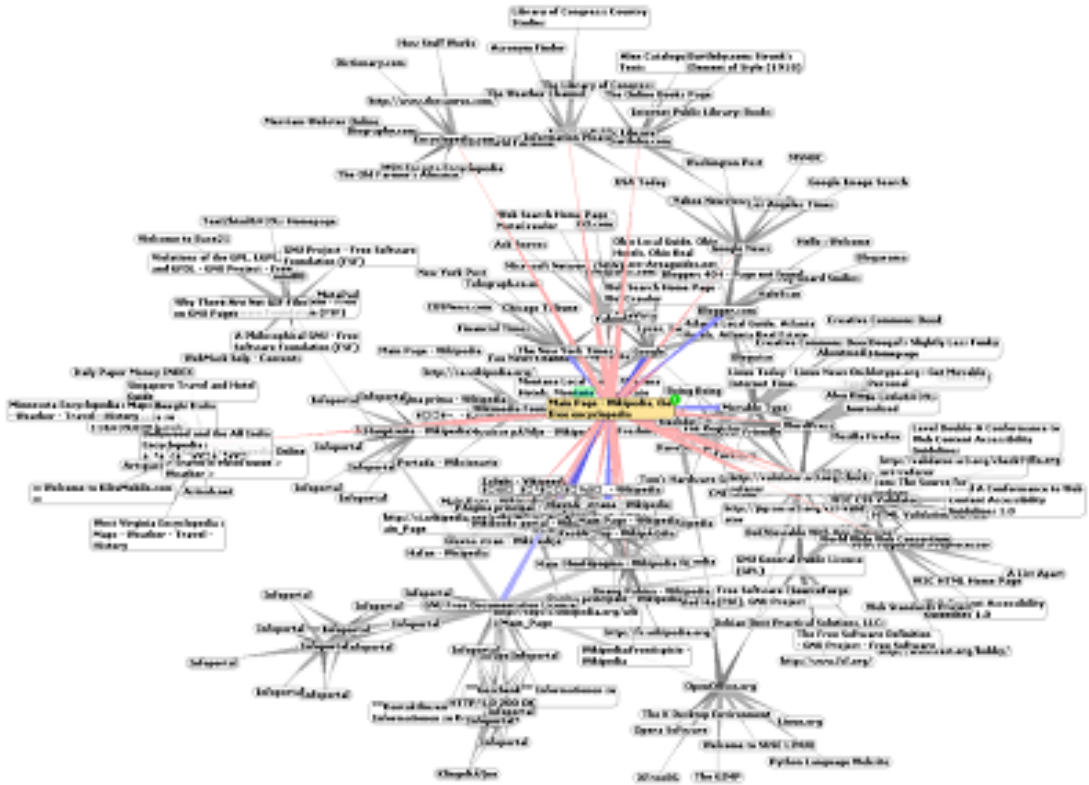
E.g., in the P2P network,

$$W_{ij}^{u} \begin{cases} > 0, & \text{if user } i \text{ and user } j \text{ downloaded the same content,} \\ = 0, & \text{otherwise.} \end{cases}$$



Similarity graph

Or hyperlinks in Wikipedia typically connect related pages.



Similarity graph

Otherwise, the similarity graph can be defined by relating attributes. One standard method to construct the weighted similarity graph is based on the Radial Basis Function (RBF)

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

where X_i is a vector of attributes for the i-th data point.

Labeling and Classification Functions

Suppose we would like to classify N data points into K classes (communities).

And assume that P data points are labelled. Denote by V_k, the set of labelled points in class $k = 1, ..., K$. Thus, $|V_1| + ... + |V_K| = P$.

Denote by D a diagonal matrix with its (i, i)-element equals to the sum of the i-th row of matrix W. 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.

Optimization Based Framework

Also define an $N \times K$ matrix F and call its columns F_{*k} classification functions.

The points are classified according to the rule

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

General ideas of the graph-based semi-supervised learning is to find classification functions so that

- on one hand they will be close to the corresponding labeling function,
- on the other hand they will change smoothly over the similarity graph.

These ideas can be expressed with the help of optimization formulations.

Optimization Based Framework

For instance, the Standard Laplacian based method (Zhou & Burges 2007) has the following optimization formulation

$$\min_F \{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|F_{i*} - F_{j*}\|^2 + \mu \sum_{i=1}^{N} d_{ii}\|F_{i*} - Y_{i*}\|^2\},$$

with μ as a regularization parameter.

Optimization Based Framework

We suggest to find the classification functions as a solution of the following more general optimization problem:

$$\min_F \{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\|{d_{ii}}^{\sigma-1}F_{i*} - {d_{jj}}^{\sigma-1}F_{j*}\|^2 + \mu\sum_{i=1}^{N}{d_{ii}}^{2\sigma-1}\|F_{i*} - Y_{i*}\|^2\} \quad (1)$$

Now we have two parameters μ and σ.

Optimization Based Framework

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

The classification functions for the generalized semi-supervised learning are given by

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

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

Optimization Based Framework

In particular cases, we have

- 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:
  $F_{*k} = \frac{\mu}{2+\mu}(I - \frac{2}{2+\mu}D^{\frac{-1}{2}}WD^{\frac{-1}{2}})^{-1}Y_{*k}$,
- if $\sigma = 0$, PageRank based method:
  $F_{*k} = \frac{\mu}{2+\mu}(I - \frac{2}{2+\mu}WD^{-1})^{-1}Y_{*k}$.

Random Walk interpretation

It is 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

$$\text{ppr}(j) = (1 - \alpha) e_j^T \left(I - \alpha D^{-1} W\right)^{-1}, \tag{3}$$

which is the Personalized PageRank vector with respect to seed node j.

Random Walk interpretation

Theorem

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

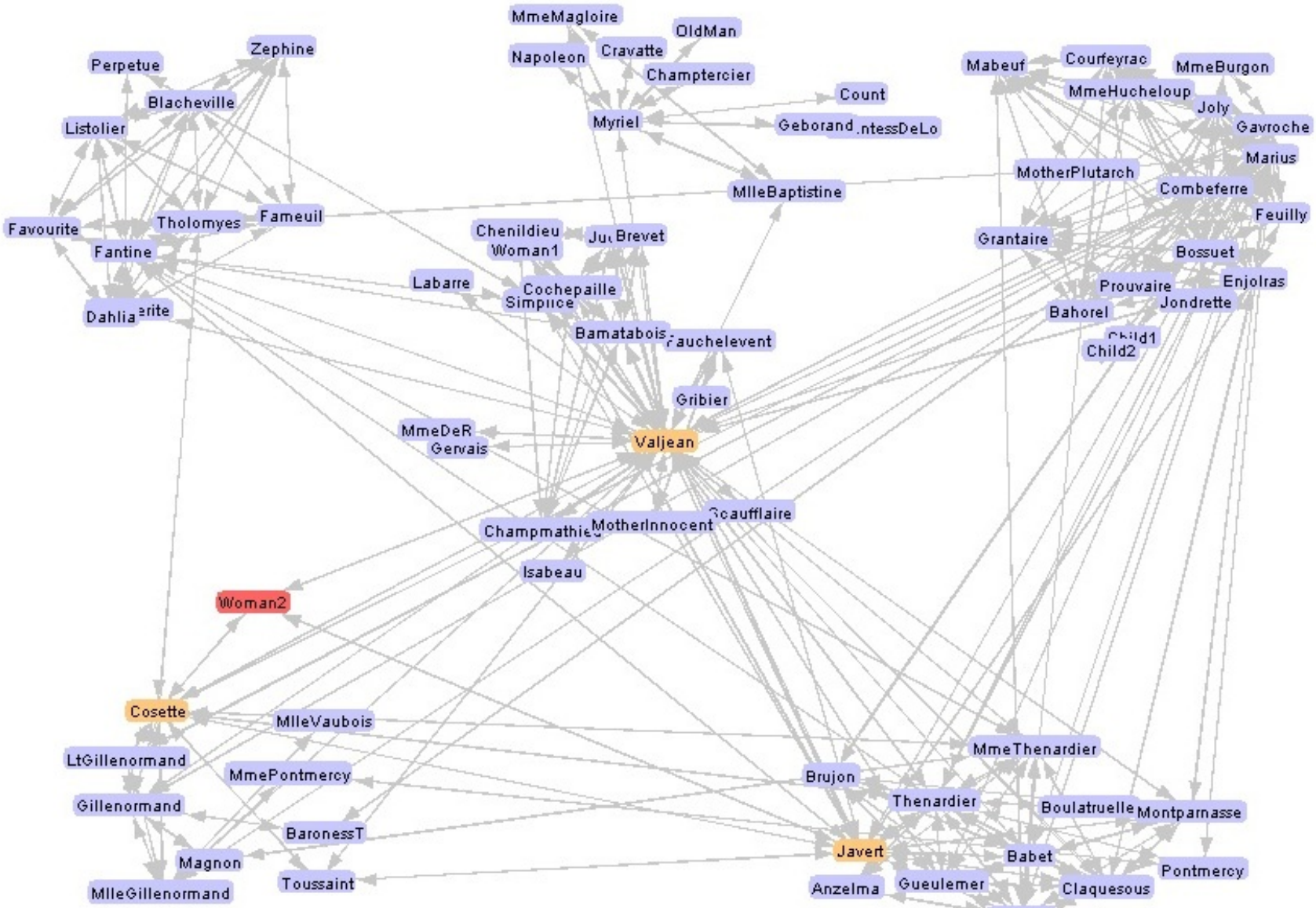
$$\sum_{p \in V_k} d_p^{\sigma} q_{pi} > \sum_{s \in V_{k'}} d_s^{\sigma} q_{si}, \quad \forall k' \neq k, \tag{4}$$

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

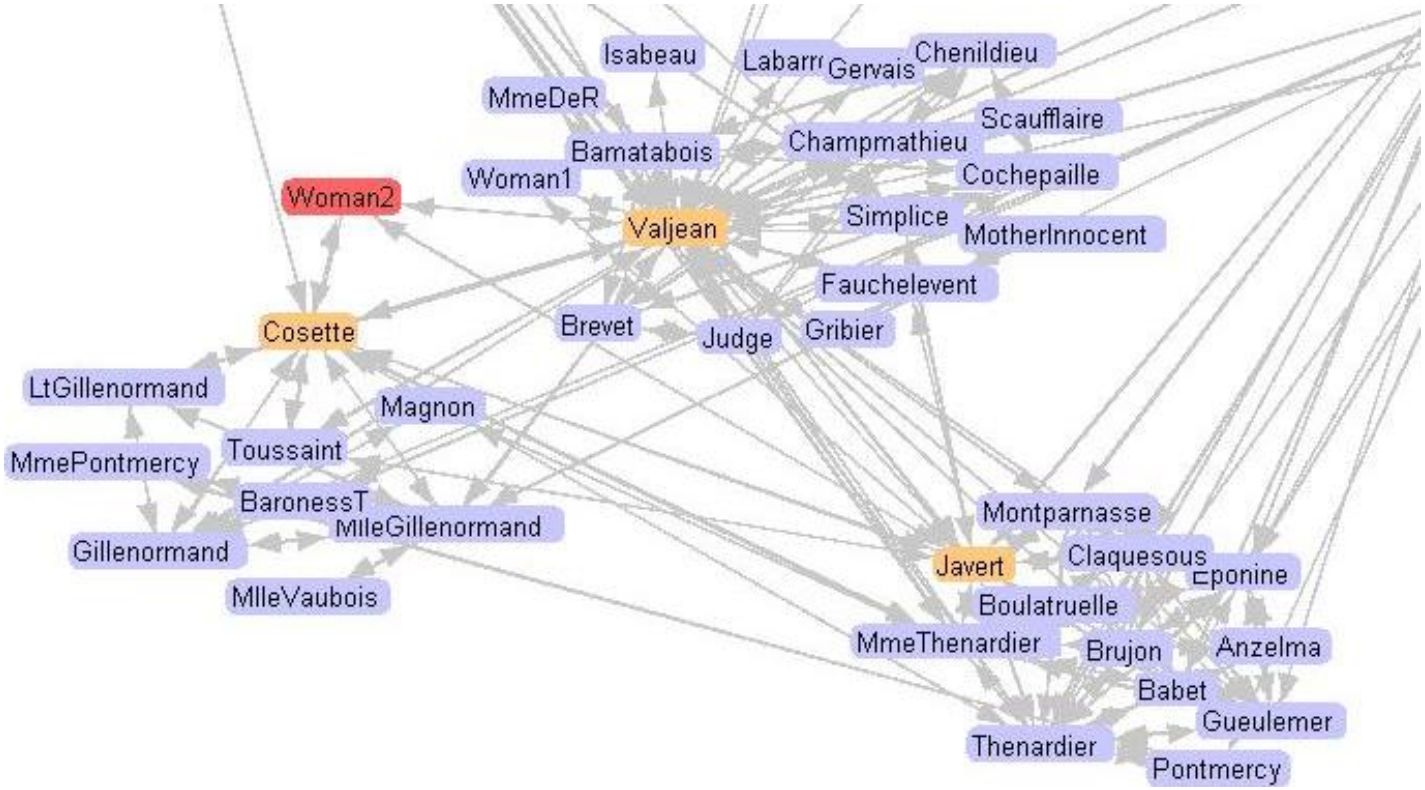
A main ingredient of the proof is the following decomposition result (K.A. & N. Litvak, 2006):

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

Random Walk interpretation



Random Walk interpretation



An interesting auxiliary result

Lemma

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

$$\frac{1}{d_j}\mathrm{ppr}_j(i) = \frac{1}{d_i}\mathrm{ppr}_i(j). \tag{5}$$

Proof: We can rewrite (3) 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.

Random Walk interpretation

The auxiliary result implies an alternative interpretation in terms of the "reversed" PageRank.

Theorem

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

$$\sum_{p \in V_k} \frac{\mathrm{ppr}_p(i)}{d_p^{1-\sigma}} > \sum_{s \in V_{k'}} \frac{\mathrm{ppr}_s(i)}{d_s^{1-\sigma}}, \quad \forall k' \neq k. \tag{6}$$

The sweeps $\mathrm{ppr}_p(i)/d_p$ introduced in (R. Andersen, F. Chung and K. Lang, 2007) now have one more application.

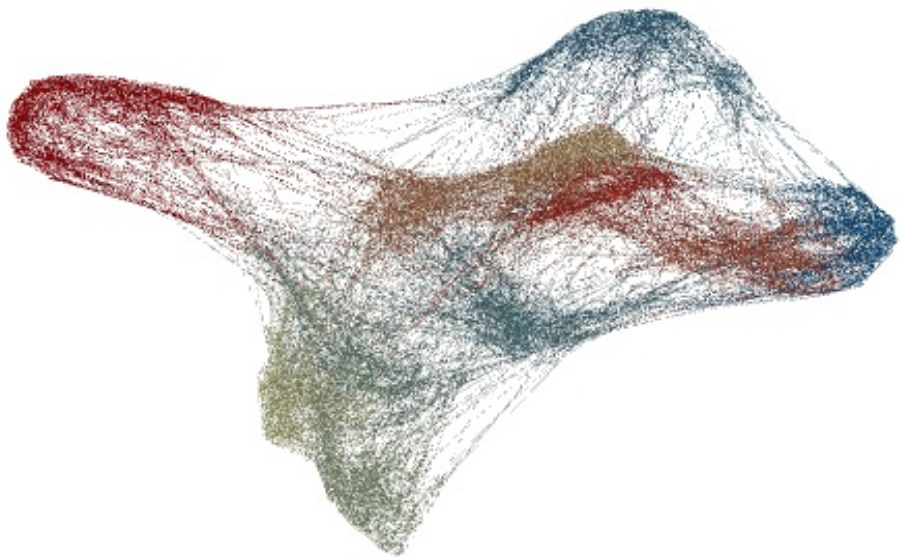
Random Walk interpretation

Theorem 1 has the following implications:

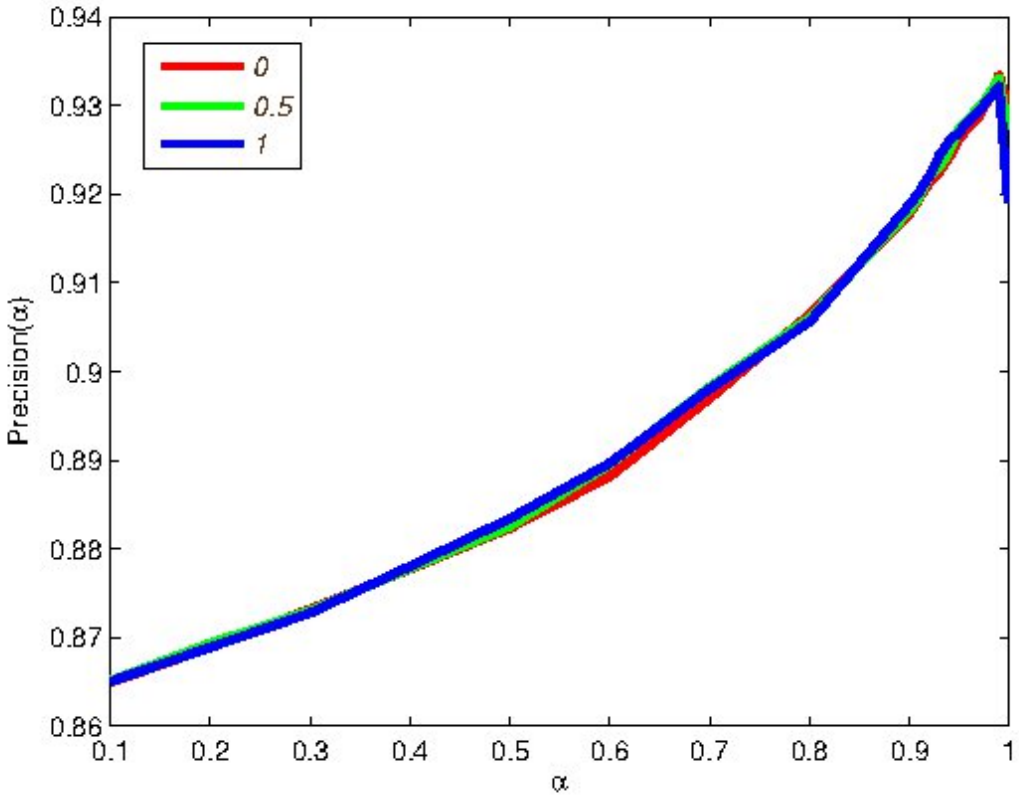
- one can decouple the effects from the choice of α and σ;
- when α goes to one, q_{pi} goes to one and classes with the largest value of $\sum_{p \in V_k} d_p^{\sigma}$ attract all points. In the case of $\sigma = 0$ and $|V_k| = const(k)$ there is a stability of classification;
- The PageRank based method attracts "border points" to a smaller class and on opposite the Standard Laplacian method attracts "border points" to a larger class.

Random Walk interpretation

- and we have one more rather surprising conclusion. Consider as an example the classification of handwritten digits (USPS dataset):



Random Walk interpretation



Random Walk interpretation

Corollary (from (4))

If the 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.*

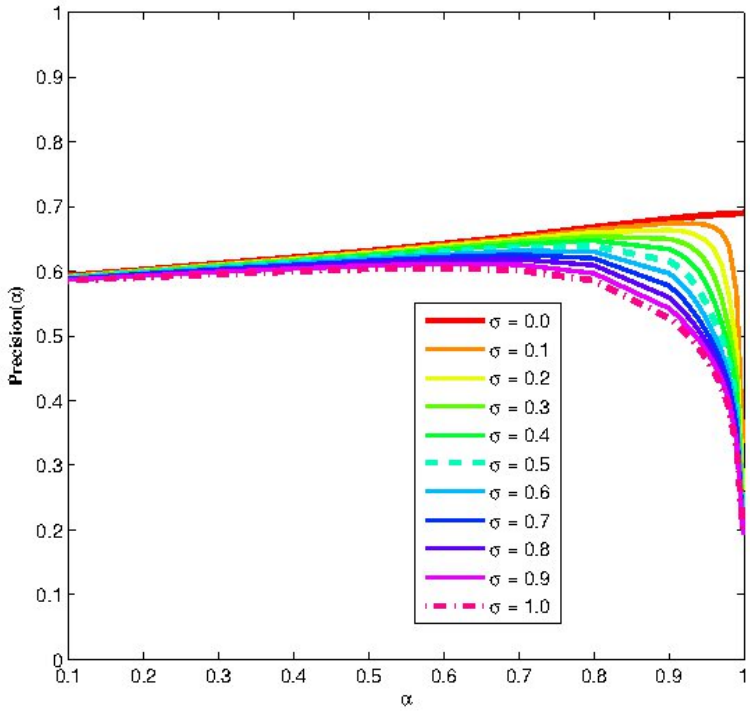
Clustered preferential attachment model

The model produced 5 unbalanced classes (1500 / 240 / 120 / 100 / 50).

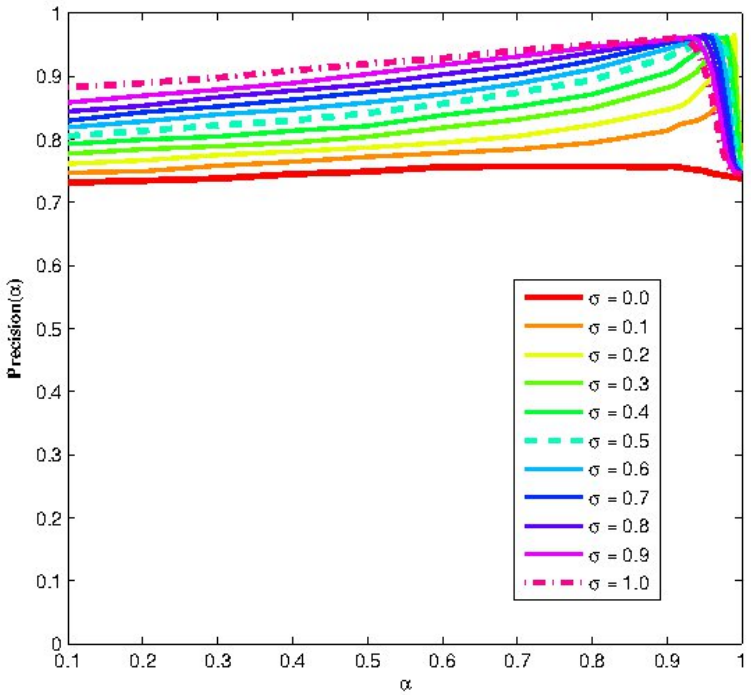
Once a node is generated, it has two links which it attaches independently with probability 0.98 within its class and with probability 0.02 outside its class.

In both cases a link is attached to a node with probability proportional to the number of existing links.

Clustered preferential attachment model



(a) Random Labelled Points



(b) Max Degree Labelled Points

Figure: Clustered Preferential Attachment Model: Precision of classification.

Robustness of the PageRank based method

Consider a dataset derived from the English language Wikipedia.

Wikipedia forms a graph whose nodes represent articles and whose edges represent hyper-text inter-article links.

We have chosen the following three mathematical topics:

- "Mathematical analysis" (MA),
- "Discrete mathematics" (DM),
- "Applied mathematics" (AM).

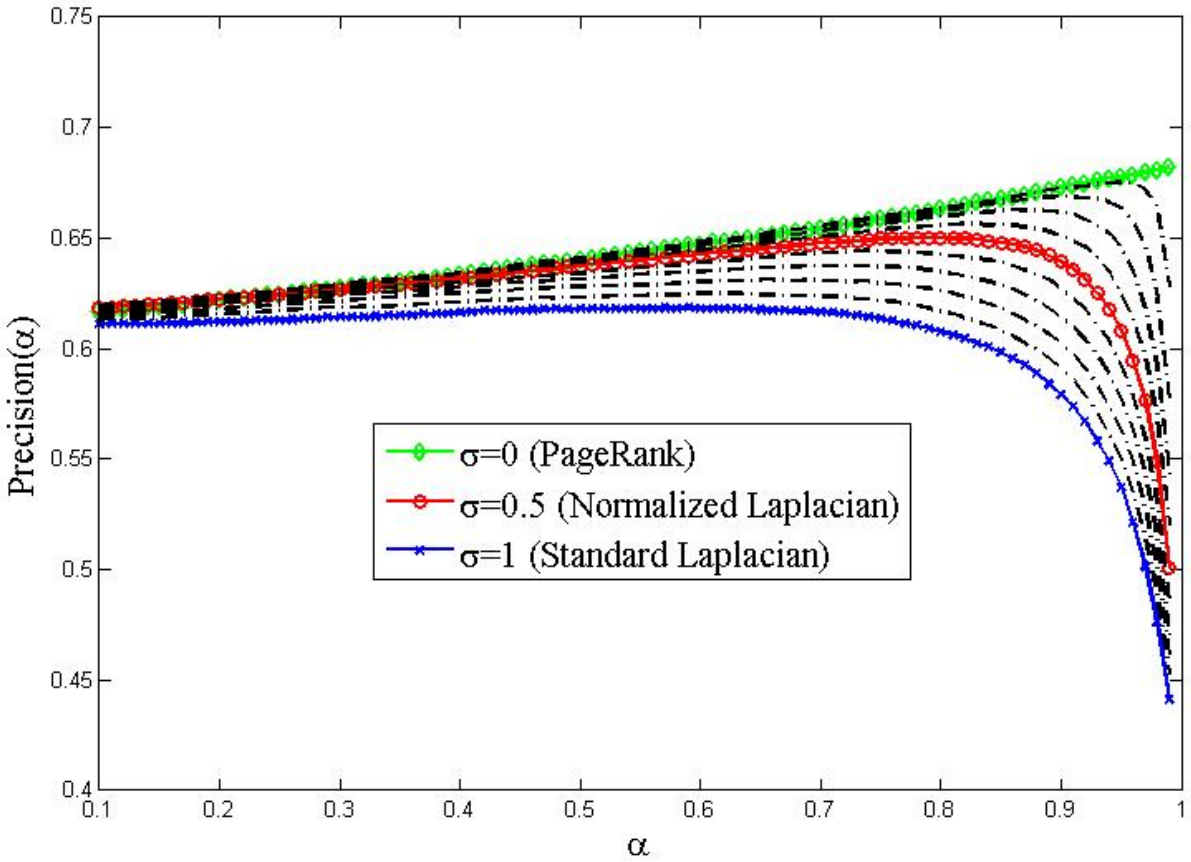
Robustness of the PageRank based method (Wikipedia example)

With the help of AMS MSC Classification 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).

Then, we have chosen uniformly at random 100 times 5 labeled nodes for each class and plotted the average precision as a function of the regularization parameter α.

Robustness of the PageRank based method (Wikipedia example)



SSL scales well (P2P classification)

Using methodology developed in Inria Planete team, we have collected several snapshots of the P2P Torrents from the whole Internet.

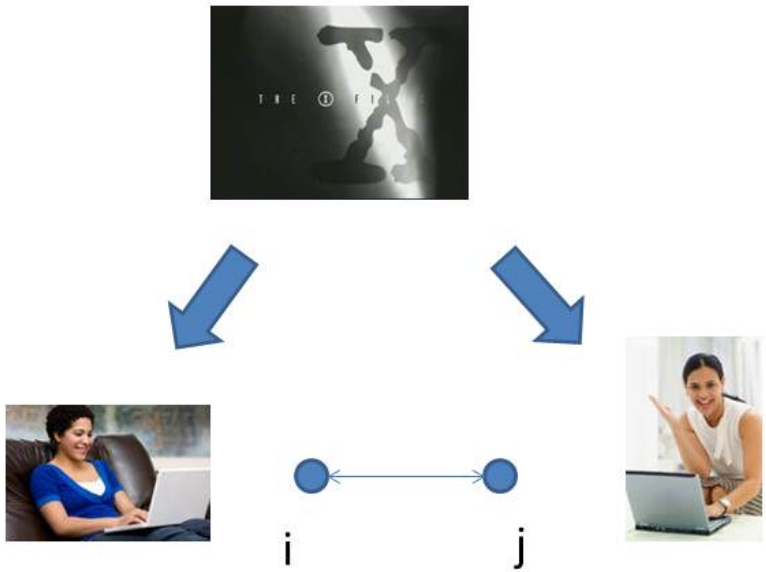
Based on these data and employing the WebGraph framework (Boldi and Vigna'04), we constructed:

- User similarity graph (1 126 670 nodes and 124 753 790 edges, after preprocessing);
- Content similarity graph (200 413 nodes and 50 726 946 edges, after preprocessing).

User similarity graph

Adjacency matrix of the user similarity graph:

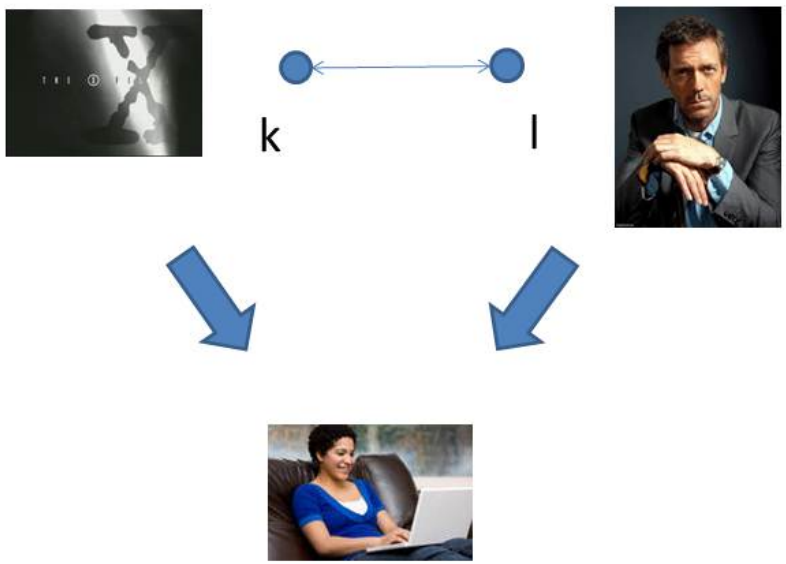
$$W_{ij}^{u} \begin{cases} > 0, & \text{if user } i \text{ and user } j \text{ downloaded the same content,} \\ = 0, & \text{otherwise.} \end{cases}$$



Content similarity graph

Adjacency matrix of the content similarity graph:

$$W^c_{kl} \begin{cases} > 0, & \text{if contents } k \text{ and } l \text{ were downloaded by at least one same user,} \\ = 0, & \text{otherwise.} \end{cases}$$



Baseline expert classification

Table: The quantity of language base line expert classifications.

Language	#content	#user
English	36465	57632
Spanish	2481	2856
French	1824	2021
Italian	2450	3694
Japanese	720	416

Table: The quantity of topic base line expert classifications.

Topic	# content	# user
Audio Music	23639	13950
Video Movies	20686	43492
TV shows	12087	27260
Porn movies	8376	7082
App. Windows	4831	2874
Games PC	4527	8707
Books Ebooks	1185	281

Classification results for the complete graphs

Using very little amount of information, we are able to classify the content and users with high accuracy.

For instance, in the dataset of 1 126 670 users, using only 50 labelled points for each language, we are able to classify the users according to their preferred language with more than 95% accuracy.

Classification of untagged content

Let us see how our method works for untagged content.

We have taken all nodes for which we have topic tags as "other video" and all edges induced by the supergraph.

(The subgraph contains 1189 nodes and 20702 edges.)

We made the expert evaluation manually by popular categories:

1. "Sport Tutorials" [ST] (116),
2. "Science Lectures" [SL] (127),
3. "Japanese Cartoons" [JC] (93),
4. "Porno" [P] (81),
5. "Software Tutorials" [SFT] (113),
6. "Movies" [M] (129).

Classification of untagged content

The cross-valiadtion matrix has a strong diagonal domination.

Classified as→	JC	M	P	SFT	SL	ST
JC	65	2	1	1	5	8
M	6	47	18	6	11	21
P	0	8	59	4	2	3
SFT	3	4	3	91	9	3
SL	5	5	3	10	85	19
ST	2	9	5	8	2	85

Table: Cross-Validation matrix for "Other Video" subgraph classification, 10 labeled points for each class, $\alpha = 0.5$.

Classification of untagged content

Curious facts:

- Most of the "other video" files with the content as "Dance Tutorials" (21 from 27) are classified into "Sport Tutorials" [ST].
- All tutorials about gun shooting (13) are classified in "Sport Tutorials", even though they have not initially been classified as "Sport Tutorials".

This automatic classification appears to be quite logical and suggests the possibility of application of graph based semi-supervised learning for refinement of P2P content and user categorization.

Unsupervised approach for choosing seeds

To choose seeds, we suggest the following empirically tested approach:

- sort nodes by their PageRank values;
- start from the top of the list;
- assign candidate nodes as seeds corresponding to the same class if they have more than 20-30% common neighbours.

Unsupervised approach for choosing seeds

Wikipedia example continued:

CLUSTER 1: Hilbert space, Partial differential equation, Functional analysis, Derivative, Banach space, Numerical analysis, Fourier transform, Lp space, Measure (mathematics), Quantum mechanics, Dirac delta function

CLUSTER 2: Combinatorics

CLUSTER 3: Dynamical system

Unsupervised approach for choosing seeds

Wikipedia example continued:
(removing first top 20 PageRank nodes)

CLUSTER 1: Analytic function, Calculus

CLUSTER 2: Tessellation

CLUSTER 3: Linear programming

References:

- D. Zhou, and C.J.C. Burges, "Spectral clustering and transductive learning with multiple views", in Proceedings of ICML 2007.
- K. Avrachenkov, V. Dobrynin, D. Nemirovsky, S.K. Pham, and E. Smirnova, "Pagerank based clustering of hypertext document collections", In Proceedings of ACM SIGIR 2008.
- K. Avrachenkov, P. Goncalves, A. Mishenin, and M. Sokol, "Generalized optimization framework for graph-based semi-supervised learning", In Proceedings of SDM 2012.
- K. Avrachenkov, P. Goncalves, and M. Sokol, "On the choice of kernel and labelled data in semi-supervised learning methods", In Proceedings of WAW 2013.

Thank you!

Any questions and suggestions are welcome.