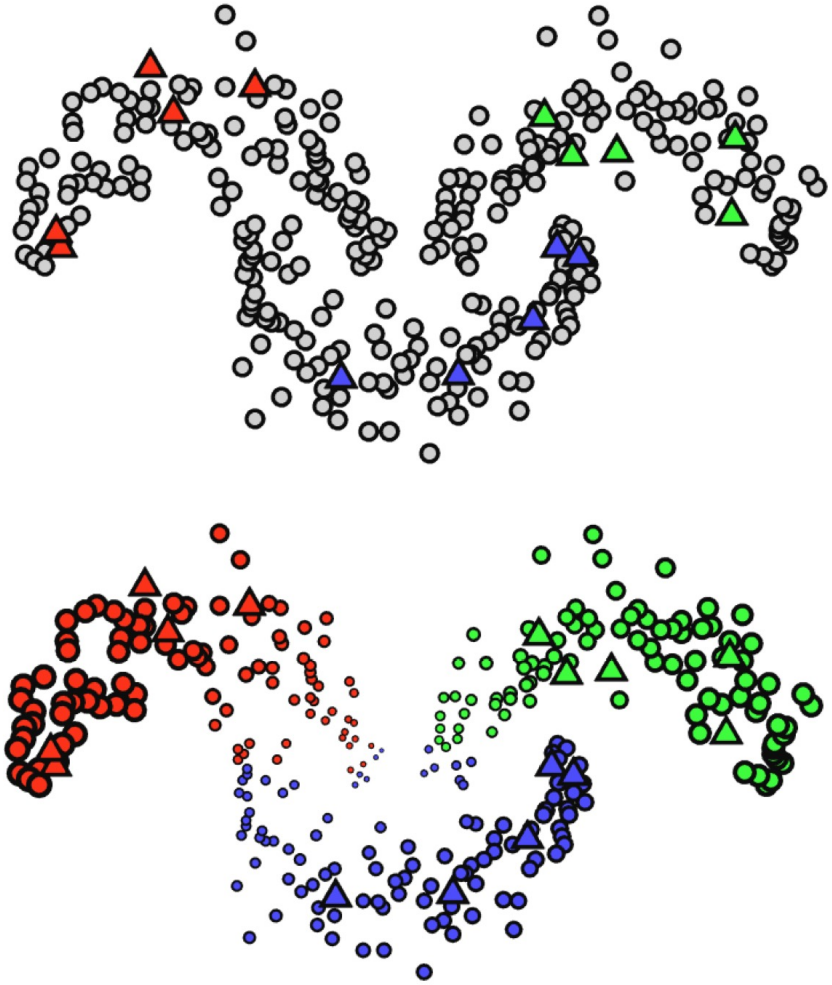
DATA SCIENCE MEI/1

University of Beira Interior
Department of Informatics

Hugo Pedro Proença,
hugomcp@di.ubi.pt, **2021/22**

Semi Supervised Learning

- **Semi-supervised learning** is obtaining increasing popularity in the literature.
- It involves both labelled and unlabeled data when training a model that makes a prediction.
- Usually, in this setting, you will have a **small amount of labeled data** and a **large amount of unlabeled data**.
- It is becoming increasingly important because it can combine **data carefully labeled by humans** with abundant unlabeled data to train (e.g.) deep neural networks.

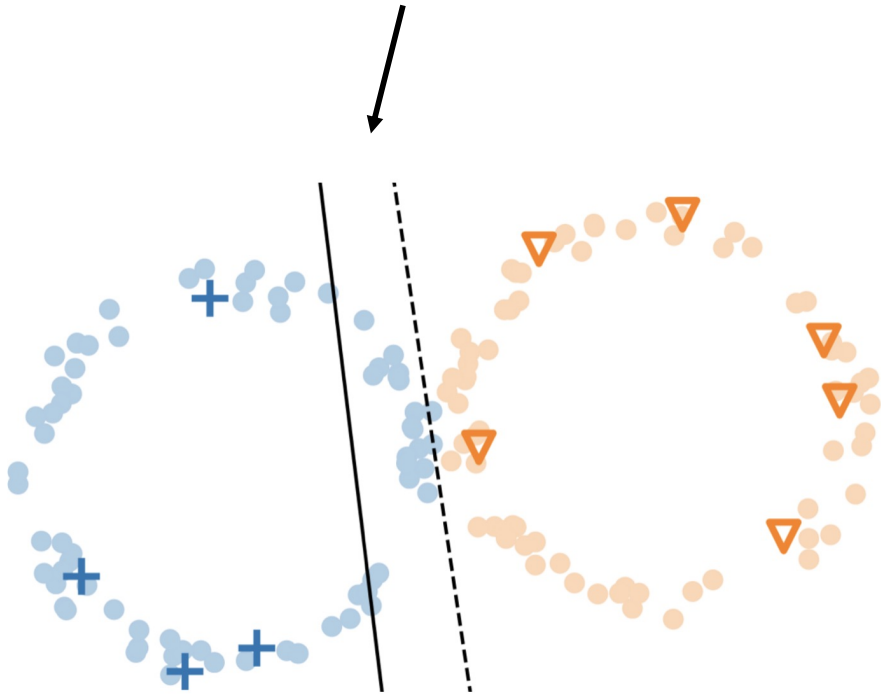


Source: https://arxiv.org/pdf/1904.04717.pdf

Semi Supervised Learning

- As an example, Visual data are available in large quantities.
- However, data reliably annotated by humans are still very scarce, in relative terms.
- Obtaining large amounts of annotated training data for every single task is not only impractical, potentially costly, but it also turns out to be error prone. The low quality of crowd-sourced annotation is a common motivation to minimize the need of annotation.
- For some problems, the results obtained by semi-supervised learning algorithms can be even better than the obtained for fully supervised models.

The continuous line represents the boundary decision plane for s fully supervised model, whereas the dashed line represents the results attained by a semi-supervised paradigm



Source: https://towardsdatascience.com/a-friendly-intro-to-semi-supervised-learning-3783c0146744

Semi Supervised Learning

- The most basic approach in semi-supervised learning is to develop an iterative process that – at each moment – **increases the number of labelled samples** used to train a **classic supervised learning model**
- The first paradigm is called **pseudo-labeling**, which uses the network itself to generate ground truth labels for the unlabeled data.
  - The model is often pretrained with the fully labeled subset, that one needs to obtain.
  - The unlabeled samples are then fed into the network and their class predictions are recorded.
    - If the largest class probability of a sample **exceeds a set threshold**, **the corresponding class is used as ground truth**.
    - This means that the pseudo-label is added to the ground-truth set, i.e., the unlabeled element starts to be used as if it was manually labeled.
  - These samples are further used to train the model in a supervised fashion. As the performance of the model gets better and better, the artificially obtained labels can **iteratively be refined using the very same technique**.

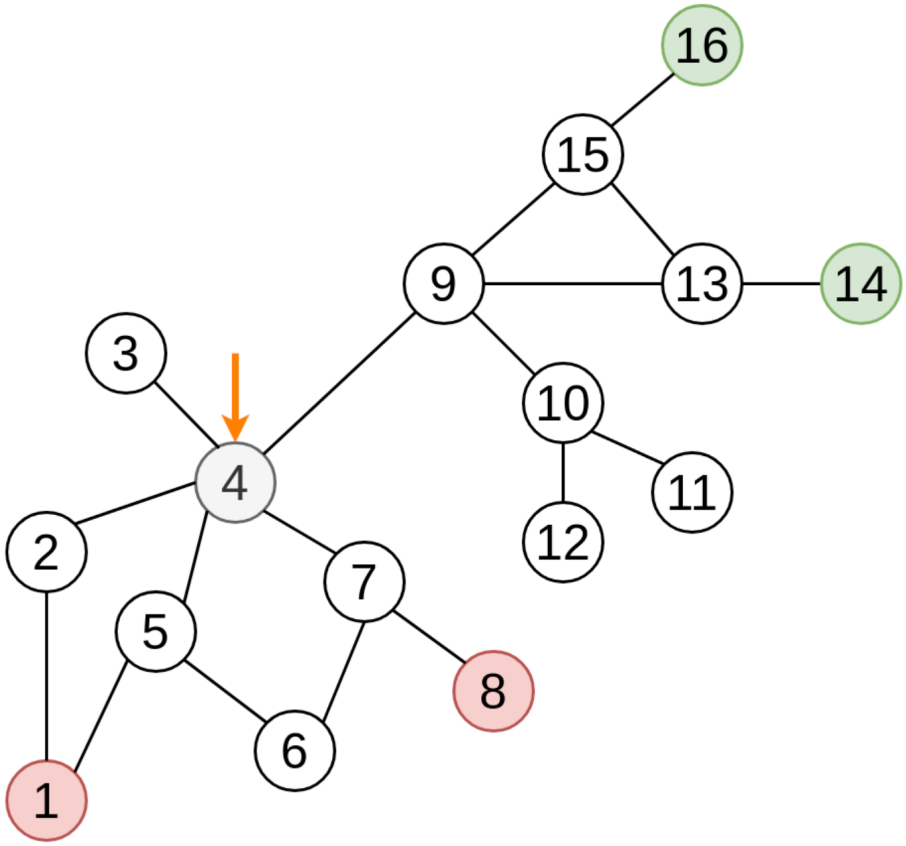
Semi Supervised Learning

- A more sophisticated technique is called "**Label Propagation Algorithm**" (LPA), and at each iteration, selects a subset of the unlabeled samples to be used in the learning set of the supervised model.
- Is an iterative algorithm where we assign labels to unlabeled points by propagating labels through the dataset. This algorithm was first proposed by Xiaojin Zhu and Zoubin Ghahramani [1] in the year 2002.
- LPA works under the transductive learning paradigm, as we want to predict labels of the unlabeled data points which are already given to us.

[1] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Technical Report. CMU-CALD-02–107, 2002.

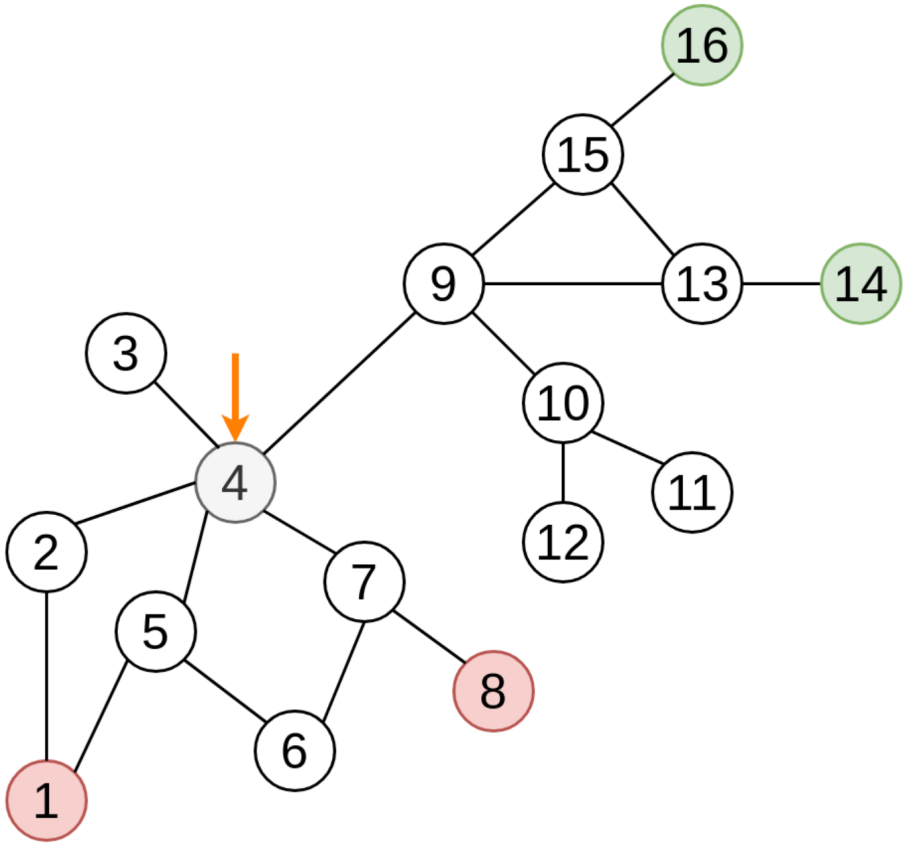
Semi Supervised Learning

- Again, the basic premise in **LPA** is that only a small set of samples is labeled, whereas for most of the data we don't have any information about the ground-truth labels.
- Suppose that we have a dataset of "n" elements (16 in the example), but where k **(k <<n)** of these are labeled (e.g., 4 out of 16 in the example).



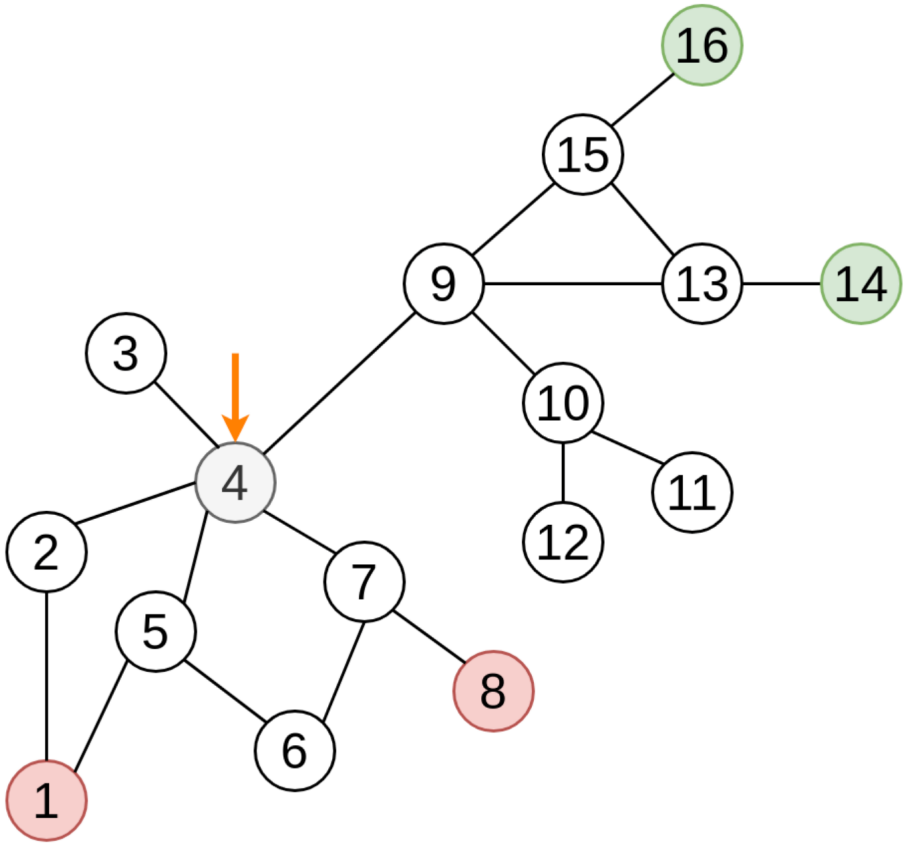
Semi Supervised Learning

- Each vertex in this graph is one instance of the dataset, while the edges represent the similarity between elements (i.e., for example the Euclidean distance).
- In such case, we can assume that elements fully dissimilar (i.e., with a distance higher than a threshold, are simply not related)



Semi Supervised Learning

- Starting from each node (vertex), we can walk for many different paths.
  - 4→3
  - 4→7→8
  - 4→9→15→16
  - …
- Every time we hit a labeled node, we stop a walk, as we reached an "**Absorbing state**"
- We can define a maximum number of (t) steps, but also set t to infinity.



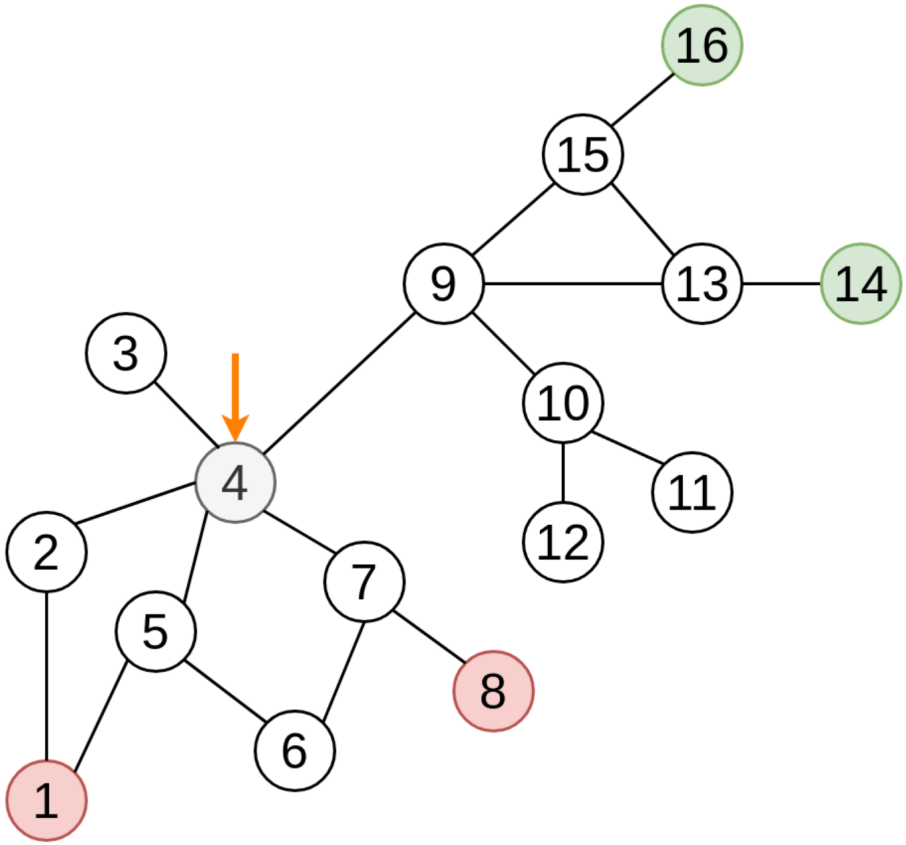
Semi Supervised Learning

- These are called “random walks”.
- The basic intuition is to label the node (4 in the example), according to the maximum number of paths (walks) that end up in an absorbing state of a specific class.

$$y_i[c] = \sum_{j \in X_l} T_{ij}{}^{t}\, y_j[c]$$

The probability of node $x_i \in X_u$ to have label c

The probability P(i→j) to jump from node x_i and end up in node x_j in t steps. We can define the number of steps to be a large number (infinity).



Semi Supervised Learning

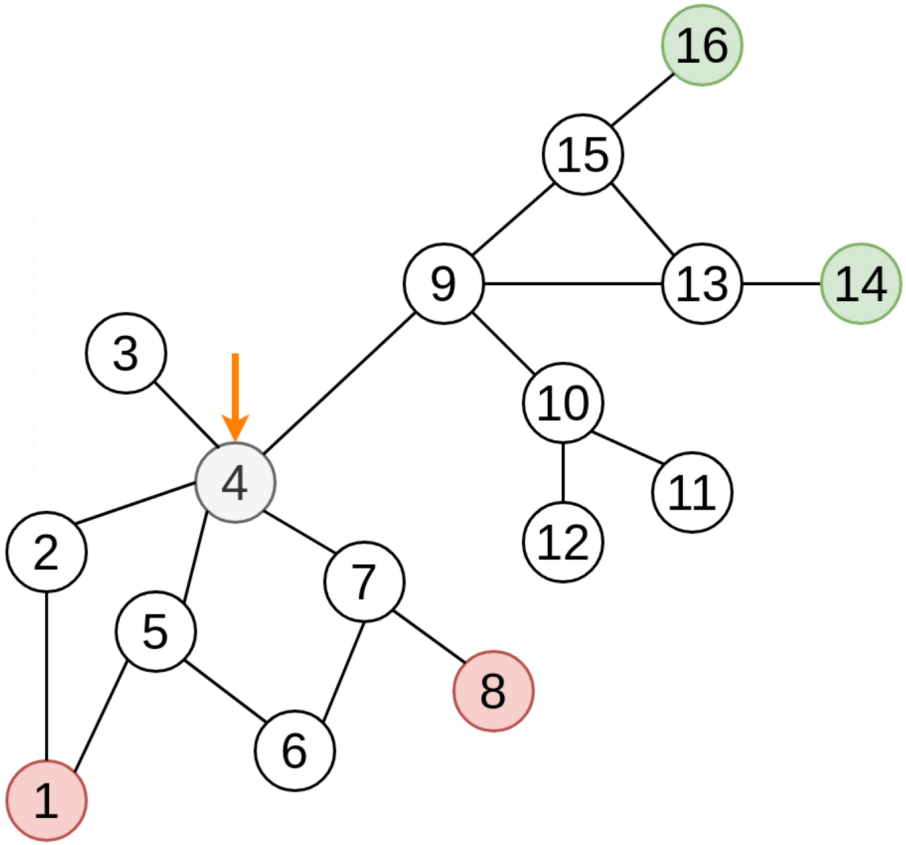
- In Matrix form, the whole process can be represented as

$$\hat{Y} = T^{t\to\infty} Y$$

Vector of labels we need to get

Matrix of probabilities to end up on different nodes

Vector of labels on those nodes

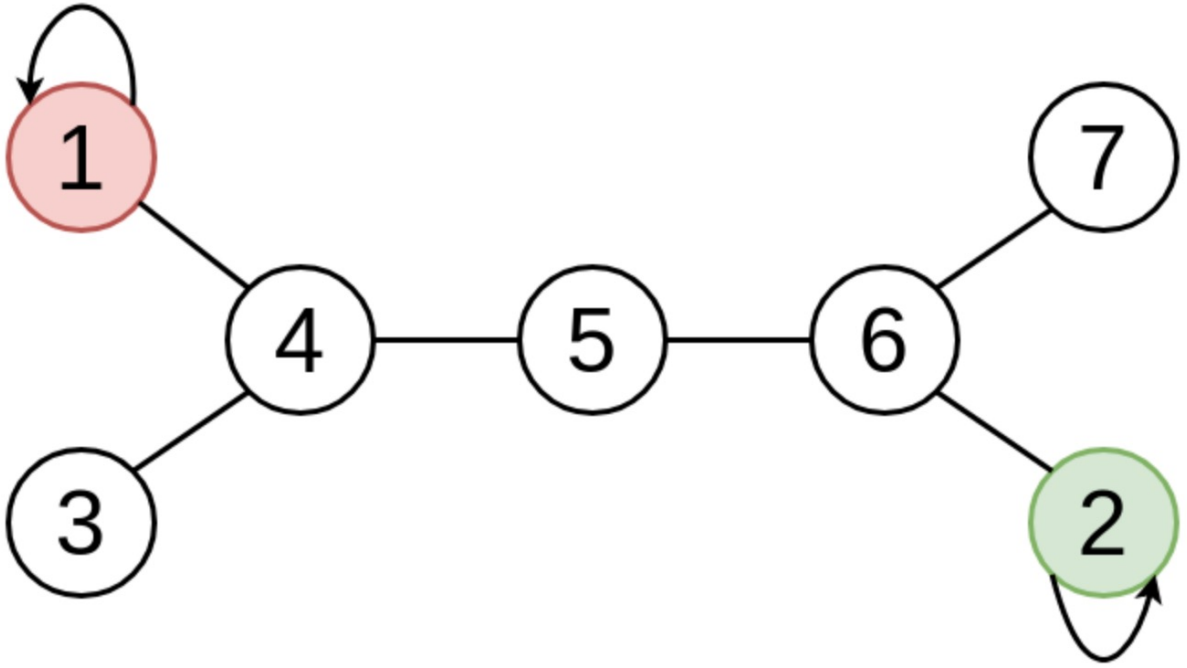


- In practice, we are interested in obtaining the labels for Y_u

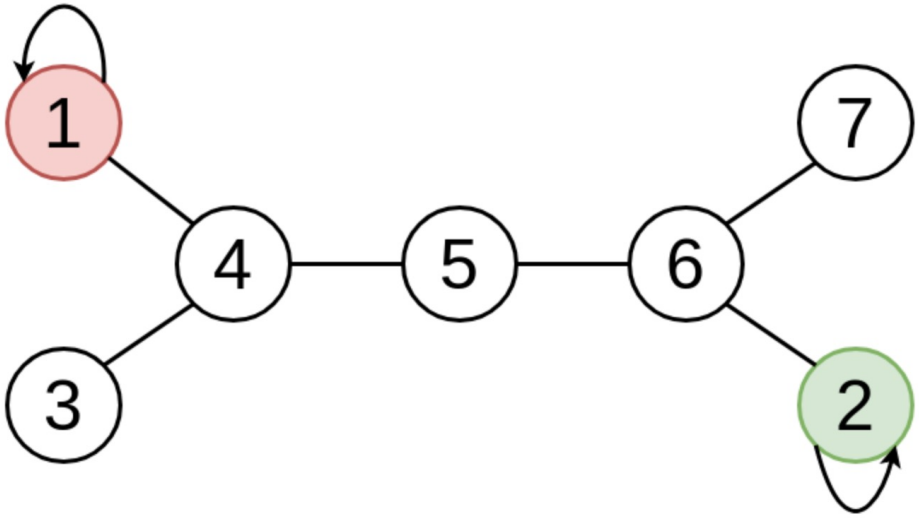
$$\begin{bmatrix} \hat{Y}_l \\ \hat{Y}_u \end{bmatrix} = T^{t\to\infty} \begin{bmatrix} Y_l \\ 0 \end{bmatrix}$$

Semi Supervised Learning

- Consider the following (simpler) example, with two absorbing nodes (1 and 2)
- For simplicity, we consider that the distances (similarity) between all elements is the same.
  - This distance can be regarded as the probability of moving between states



Semi Supervised Learning



- The corresponding transition matrix is as follows:

$$T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0.33 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0.33 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

D=Degree Matrix

A=Adjacency Matrix

$T = D^{-1}. A$

Semi Supervised Learning

- The matrix **T can be divided into 4 parts:**
  - T_{ll} — Probability to get from labelled nodes to labelled nodes
  - T_{lu} — Probability to get from labelled nodes to unlabelled nodes
  - T_{ul} — Probability to get from unlabelled nodes to labelled nodes
  - T_{uu} — Probability to get from unlabelled nodes to unlabelled nodes

$T =$

	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	0	1	0	0	0
4	0.33	0	0.33	0	0.33	0	0
5	0	0	0	0.5	0	0.5	0
6	0	0.33	0	0	0.33	0	0.33
7	0	0	0	0	0	1	0

$$T = \begin{bmatrix} T_{ll} & T_{lu} \\ T_{ul} & T_{uu} \end{bmatrix} = \begin{bmatrix} I & 0 \\ T_{ul} & T_{uu} \end{bmatrix}$$

Semi Supervised Learning

- The final transition matrix is obtained by multiplying T by itself "t" times ("t→∞").
  - In practice, we multiply the matrix up to the moment when its values do not change more than a threshold anymore.
- The final matrix will have the form:

$$= \begin{bmatrix} I & 0 \\ (I - T_{uu})^{-1}.T_{ul} & 0 \end{bmatrix}$$

Important part: Gives the probability of moving from one unlabeled to a labeled sample

Semi Supervised Learning

- The probability of reaching a labeled state (element) from an unlabeled one is given by:

$$\hat{Y}_u = (I - T_{uu})^{-1}.T_{ul} \times Y_l$$

$$\hat{Y}_u = \begin{matrix} 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix}\begin{bmatrix} 0.7316 & 0.2390 \\ 0.7316 & 0.2390 \\ 0.4853 & 0.4853 \\ 0.2390 & 0.7316 \\ 0.2390 & 0.7316 \end{bmatrix} \times \begin{matrix} & \text{R} & \text{G} \\ 1 & 1 & 0 \\ 2 & 0 & 1 \end{matrix} = \begin{matrix} & \text{R} & \text{G} \\ 3 & 0.7316 & 0.2390 \\ 4 & 0.7316 & 0.2390 \\ 5 & 0.4853 & 0.4853 \\ 6 & 0.2390 & 0.7316 \\ 7 & 0.2390 & 0.7316 \end{matrix}$$

- Then, we consider the set of pseudo-labels that are higher than a certain threshold.
- In the given example, supposing that the threshold = 0.5, the “3” and “4” elements will assume the “**red**” label, while “6”/”7” will assume the “**green**” one.

Semi Supervised Learning - Exercise

- Consider the following partially labeled dataset. Implement a "Python" script to obtain the next generation of labeled samples, according to the "LPA" algorithm.

X1	X2	X3	X4	Y
1	0.5	0.3	0.4	1
0.5	0.1	0.2	0.7	?
0.2	0.3	0.3	0.6	?
0.7	0.1	0.8	0.9	0
0.3	0.9	0.7	0.1	?
0.2	0.5	0.8	0.2	0
0.3	0.4	0.1	0.2	?
0.1	0.1	0.1	0.3	?
0.2	0.9	0.8	0.8	1