13 pages (including this one).

- The theory part of the homework is due Friday, October 23 by 5:00 P.M.
- The experimental part of the homework is due Friday, October 30 by 5:00 P.M.
- Your proofs should be neat and legible. You are encouraged to use LaTeX or Word.
- You may use any textbook and the notes you have taken in class.
- If you refer to a result from class, make sure you clearly state the entire result. If you use a textbook, please refer to the result from the book by number (e.g., Theorem 17.20 on page 187, of Matrix Analysis and Applied Linear Algebra).
- You will need to use MATLAB for the experimental part of the homework.

Theory		
Problem	Points	Score
1	5	
2	5	
3	5	
4	5	
5	5	
6	10	
7	10	
8	10	
9	5	
10	20	
11	10	
12	10	
Total	100	

Experiments		
Problem	Points	Score
13	10	
14	10	
15	10	
16	30	
17	30	
18	5	
19	5	
Total	100	

Theory

Introduction

Social networks (e.g., facebook, LinkedIn, etc) can be modeled using unweighted undirected graphs. Formally, we model the social network composed of n individuals using an $n \times n$ symmetric adjacency matrix, $\boldsymbol{A}$, defined by

$$a_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked, or friends,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For such a network, we can define the notion of *community* (e.g., LinkedIn has a notion of "groups" for alumni of a given university, or people interested in a professional activity).

Informally, a community is a cluster of nodes (vertices) that are more connected with each other, than with the rest of the network. While the notion of community does not have a rigorous mathematical definition, simple mathematical models can be constructed and analyzed. The goal of this test is to study theoretically and experimentally such a model.

The *planted partition model* (also known as the *block stochastic model*) with two communities of equal size provides a simple example of a mathematical model of two communities. Much research has been expanded on the theoretical understanding of this model in the last five years (some conjectures were just recently proved this year).

The planted partition $G(n, p, q)$ is the set of symmetric random matrices $A = (a_{i,j})$ defined as follows.

Definition 1. *We consider a set of n nodes (without loss of generality, we choose n to be even). We randomly divide the nodes into two sets of equal size. Each set represents one community: community 1, and community 2.*

For two nodes i and j, with $i \leqslant j$. we randomly connect the nodes with a probability p if i and j are part of the same community. Otherwise, the nodes i and j are randomly connected with a probability q. We choose $0 \leqslant q < p \leqslant 1$.

Finally, $a_{j,i} = a_{i,j}$ (undirected graph).

We note that we choose $p > q$, and therefore the network is *assortative*: people in the same community are more linked than people from different communities.

Given a random realization A of a planted partition $G(n, p, q)$, the goal is to identify the two communities. Clearly, if $p = 1$ and $q = 0$, the problem is simple. Our ability to detect the partition will decrease as $p - q$ goes to zero.

Any adjacency matrix A of the planted partition $G(n,p,q)$ model can be generated using the following model,

$$A = T\mathcal{B}T^T \tag{2}$$

where

$$\mathcal{B} = \begin{bmatrix} B_{1,1}(p) & \cdots & B_{1,n/2}(p) & B_{1,n/2+1}(q) & \cdots & B_{1,n}(q) \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ B_{n/2,1}(p) & \cdots & B_{n/2,n/2}(p) & B_{n/2,n/2+1}(q) & \cdots & B_{n/2,n}(q) \\ B_{n/2,1}(q) & \cdots & B_{n/2,n/2}(q) & B_{n/2,n/2+1}(p) & \cdots & B_{n/2,n}(p) \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ B_{n,1}(q) & \cdots & B_{n,n/2}(q) & B_{n,n/2+1}(p) & \cdots & B_{n,n}(p) \end{bmatrix} \tag{3}$$

where all the $B_{i,j}(p)$ and $B_{k,l}(q)$ are independent Bernoulli random variables, and T is a random permutation matrix.

We recall that a permutation σ on $\{1,\ldots,n\}$ is a one-to-one map from $\{1,\ldots,n\}$ to $\{1,\ldots,n\}$. Given a permutation σ, we define the permutation matrix T as follows:

$$t_{i,j} = \begin{cases} 1 & \text{if } j = \sigma(i) \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

T has one and only one entry equal to 1 in each row and in each column; all the other entries are zero. T performs the random assignment of the nodes to the two communities, while the matrix $\mathcal{B}$ encodes the presence of random edges in each community.

The decomposition (2) should be interpreted as follows:

(i) Randomly generate the matrix of edges given by $\mathcal{B}$, where the first community corresponds to the nodes 1 to $n/2$ and the second community corresponds to the nodes $n/2+1$ to n.

(ii) Randomly re-index the row indices using the permutation T, and the column indices using T^T. Effectively, these combined steps re-index the nodes of the graph.

We make the following assumption: we assume that we have an oracle that gives us access to T. As we will see, our algorithm will not require this assumption, but using this assumption simplifies greatly the analysis of the algorithm.

> In summary, in the remainder of the theoretical analysis, we assume that T is the identity matrix. We therefore assume that the the first community corresponds to the nodes 1 to $n/2$, and the second community corresponds to the nodes $n/2+1$ to n.
>
> Our goal is to devise an algorithm that discovers these two communities, but does not require the (secret) information about the location of the communities.

Our approach relies on the computation of the second dominant eigenvector of A. In the following, you will derive some approximations to the eigenvalues and eigenvectors of A in order to construct an algorithm to detect the two communities.

The eigenvectors of the expected value of A

1. We consider realizations A of the planted partition model $G(n,p,q)$. Prove that the expected adjacency matrix $M = \mathbb{E}[A]$, computed over all possible realizations of $\mathcal{B}$ in (3), is given by

$$M = \begin{bmatrix} p & \cdots & p & q & \cdots & q \\ \vdots & & \vdots & \vdots & & \vdots \\ p & \cdots & p & q & \cdots & q \\ q & \cdots & q & p & \cdots & p \\ \vdots & & \vdots & \vdots & & \vdots \\ q & \cdots & q & p & \cdots & p \end{bmatrix} \tag{5}$$

   [Hint: the expected value of a Bernoulli random variable with probability success p, $B(p)$, is $\mathbb{E}[B(p)] = 1 \times p + 0 \times (1-p) = p$].

2. The degree matrix is defined as the diagonal matrix with entries $d_i = \sum_{j=1}^{n} A_{i,j}$. Derive the expression of the expected degree matrix, $\mathbb{E}[D]$ in terms of p, and q.

3. Prove that the vector $w_1 = \frac{1}{\sqrt{n}}\mathbb{1}$, where $\mathbb{1}_i = 1, i = 1, \ldots, n$, is an eigenvector of M. Compute the corresponding eigenvalue, μ_1.

We now construct all the other eigenvectors of M. We first introduce the vector w_2 defined by

$$w_2(i) = \frac{1}{\sqrt{n}} \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant n/2 \\ -1 & \text{otherwise} \end{cases} \tag{6}$$

4. Prove that w_2 is an eigenvector of M. Compute the corresponding eigenvalue, μ_2.

We now construct a family of other eigenvectors: they will be square wave functions that oscillates more and more. This family is called the Walsh functions, and can be defined by induction. We will only consider the first four basis elements, w_1, w_2, w_3 and w_4. We already defined w_1 and w_2. We now define w_3 and w_4. Let

$$w_3(i) = \frac{1}{\sqrt{n}} \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant n/4, \\ -1 & \text{if } n/4 < i \leqslant 3n/4, \\ 1 & \text{if } 3n/4 < i \leqslant n, \end{cases} \tag{7}$$

and

$$w_4(i) = \frac{1}{\sqrt{n}} \begin{cases} 1 & \text{if } 1 \leqslant i \leqslant n/4, \\ -1 & \text{if } n/4 < i \leqslant n/2, \\ 1 & \text{if } n/2 < i \leqslant 3n/4, \\ -1 & \text{if } 3n/4 < i \leqslant n. \end{cases} \tag{8}$$

5. Sketch the graph of w_3, and w_4.

6. Prove that w_3 and w_4 are in the null space of the matrix $\mathbb{E}[A]$. In other words, $Mw_n = 0, n = 3, 4$

7. Prove that

$$M = \mu_1 w_1 w_1^T + \mu_2 w_2 w_2^T \tag{9}$$

8. Describe a simple algorithm to recover the two communities using the eigenvectors of M.

To further study the spectrum of A, and the corresponding eigenvectors, we decompose each matrix A, generated by (2) and (3), as follows,

$$A = M + X. \tag{10}$$

Where the matrix X is a symmetric random matrix. The upper triangular part of the random matrix X is defined as follows:

if $1 \leqslant i \leqslant j \leqslant n/2$, or $n/2 < i \leqslant j \leqslant n$, then

$$x_{i,j} = \begin{cases} 1-p & \text{with probability } p, \\ -p & \text{with probability } 1-p, \end{cases} \tag{11}$$

and if $1 \leqslant i \leqslant n/2$, and $n/2 < j \leqslant n$, then

$$x_{i,j} = \begin{cases} 1-q & \text{with probability } q, \\ -q & \text{with probability } 1-q. \end{cases} \tag{12}$$

Finally, the matrix X is made symmetric,

$$x_{i,j} = x_{j,i}, \quad 1 \leqslant i < j \leqslant n. \tag{13}$$

9. Prove that $\mathbb{E}[X] = 0$, where the expectation is computed over all possible realizations of the matrix $\mathcal{B}$.

Separating the dominant eigenvalues from the bulk

X is a symmetric random matrix with independent entries that have mean zero. We consider the (real) eigenvalues of X, and the corresponding (non normalized) empirical spectral distribution (ESD for short) defined by

$$\mu([a,b]) = \frac{1}{n} \# \{i : \lambda_i \in [a,b]\}. \tag{14}$$

One can show (see homework 1), that the empirical spectral distribution converges toward a slightly modified form of the Wigner semi-circle law, given by

$$\frac{1}{\pi(p+q)} \sqrt{2n(p+q) - \lambda^2}. \tag{15}$$

The decomposition of A given by (10) makes it possible to estimate the spectrum of A for large n.

- The dominant eigenvalue is approximately equal to μ_1, and is given by

$$\lambda_1 = \frac{n(p+q)}{2} + 1. \tag{16}$$

  The corresponding eigenvector is approximately equal to w_1.

- The second largest eigenvalue is approximately equal to μ_2, and is given by

$$\lambda_2 = \frac{n(p-q)}{2} + \frac{p+q}{p-q}. \tag{17}$$

  The corresponding eigenvector is approximately equal to w_2.

- The remaining eigenvalues are given by the semi-circle law described by equation (15).

Algorithm `Partition`

* compute the second dominant eigenvector, v_2, of A, associated with the second largest eigenvalue λ_2.

* for $i = 1$ to n
  if the coordinate i of v_2 is positive,$(v_2)_i > 0$, then
  assign node i to community 1
  else
  assign node i to community 2.
  end
  end

Figure 1: Detection of the planted partitions using the second dominant eigenvector.

The algorithm `Partition`

We propose the algorithm `Partition` described in Fig. 1 to detect the two communities.

10. Prove that if
$$n(p-q) > \sqrt{2n(p+q)}, \tag{18}$$
then λ_2 can be separated from the continuous "semi-circle" bulk, and therefore the two communities can be detected. If this conditions fails, then the second eigenvector v_2 of A is part of the eigenvectors associated with the semi-circle, and the algorithm fails to detect the communities.

    This condition can be understood as a requirement about the contrast between the density of connections within each community in comparison to the density of connections between communities.

11. We consider the regime where the two communities are fully connected, and we choose
$$p = \alpha\frac{\log(n)}{n}, \qquad \text{and} \qquad q = \beta\frac{\log(n)}{n}. \tag{19}$$
Prove that condition (18), which guarantees that the algorithm `Partition` will succeed, takes the form
$$\alpha - \beta > \frac{2}{\sqrt{\log n}}\sqrt{\frac{\alpha+\beta}{2}}. \tag{20}$$

12. In practice, the average degree of a node in a social network is much smaller than $O(\log(n)/n)$, and is better modeled using the following values for p and q,
$$p = \frac{a}{n}, \qquad \text{and} \qquad q = \frac{b}{n}. \tag{21}$$

We note that these choices of p and q do not guarantee that the entire graph is connected. Prove that condition (18), which guarantees that the algorithm `Partition` will succeed, is now equivalent to

$$\frac{a-b}{2} > \sqrt{\frac{a+b}{2}}. \tag{22}$$

Experiments

In this part, we evaluate the performance of the algorithm `Partition` using the planted partition model. We use a random permutation T to randomly assign the nodes to the two partitions, as described in (2).

The planted partitions and the algorithm `Partition` in MATLAB

13. Given p, q, and n, write a MATLAB function that constructs the adjacency matrix A of the planted partition model, with two partitions of size $n/2$. You will use a random permutation matrix as described in equation (2) to randomly partition the nodes.

The following snippets of code may be useful:

- Fig. 2 constructs a random permutation matrix T.
- Fig. 3 constructs a planted partition where community 1 corresponds to the first $n/2$ nodes, and community 2 corresponds to the last $n/2$ nodes.

```
I = eye(n);
ix = randperm (n);
T = I(ix,:);
```

Figure 2: Generation of a random permutation matrix `T`.

```
n2 = n/2;

P   = random('bino', 1, p, n2, n2); % upper left block
dP2 = random('bino', 1, p, n2, 1); % diagonal of the lower right block
Q   = random('bino', 1, q, n2, n2); % upper right block

% carve the two triangular and diagonal matrices that we need
U = triu(P, 1);
L = tril(P,-1);
dP = diag(P);

A0 = U + U' + diag(dP);
A1 = Q;
A2 = Q';
A3 = L + L' + diag(dP2);

A =[A0 A1;A2 A3];
```

Figure 3: Generation of the adjacency matrix `A` of a planted partition defined over n nodes.

14. Implement the algorithm `Partition` defined in Fig. 1.

15. Let ω be the indicator set of the true partitions,

$$\omega_i = \begin{cases} 1 & \text{if } i \text{ belongs to partition 1,} \\ -1 & \text{if } i \text{ belongs to partition 2.} \end{cases} \tag{23}$$

Let $\widetilde{\omega}$ be the indicator set of the partition estimated using the algorithm `Partition`,

$$\widetilde{\omega}_i = \begin{cases} 1 & \text{if } (v_2)_i > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{24}$$

We define the notion of overlap between the true partition ω and the recovered partition $\widetilde{\omega}$. Because partitions 1 and 2 play symmetrical role, we measure the succes by computing the maximum correlation between ω and either $\widetilde{\omega}$, or $-\widetilde{\omega}$, and we define the unnormalized overlap as

$$\texttt{rawoverlap} = \max\left(\sum_{1=1}^{n} \delta_{\omega_i,\,\widetilde{\omega}_i},\quad \sum_{1=1}^{n} \delta_{-\omega_i,\widetilde{\omega}_i}\right) \tag{25}$$

where the Kronecker symbol is defined by

$$\delta_{\omega_i,\,\widetilde{\omega}_i} = \begin{cases} 1 & \text{if } \omega_i = \widetilde{\omega}_i \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Because a random choice for $\widetilde{\omega}$ returns a non zero `rawoverlap`, we normalize the overlap function and define,

$$\texttt{overlap} = \frac{2}{n}\texttt{rawoverlap} - 1. \tag{27}$$

Compute the value of `overlap` when $\widetilde{\omega} = \omega$. Prove that a random guess for the detection of the communities (flip a coin for each node in order to determine the community it belongs to) returns `overlap` = 0.

In the following we admit that `overlap` $\geqslant 0$ for the values of p and q that we use. The quantity `overlap` can then be interpreted as a probability of successfully detecting the communities.

Dense communities

We consider the regime where

$$p = \alpha\frac{\log(n)}{n}, \qquad \text{and} \qquad q = \beta\frac{\log(n)}{n}. \tag{28}$$

16. For $n = 300$, and for each combination of $\alpha = 5, 6, \ldots, 50$ and $\beta = 1, 2, \ldots, 50$ generate 20 random realizations of the planted partition model. Use the algorithm `Partition` to detect the communities, and compute the overlap score `overlap`.

    You will represent the results of this experiment by constructing a matrix `score`(α, β) that will contain the mean (computed over the 20 experiments) of the `overlap` score.

    Display the matrix in grayscale using the MATLAB function `imagesc`, and overlay the implicit curve defined by (20) that defines the condition on α on β for the detectability of the communities (see Fig. 6-left for an example of the figure).

Sparse communities

We now consider the regime of sparse communities where,

$$p = \frac{a}{n} \qquad \text{and} \qquad q = \frac{b}{n}. \tag{29}$$

Figure 4 displays an example of adjacency matrix A in the case $a = 10, b = 3$, and $n = 300$. Note that in such a sparse regime, each community is usually disconnected: it is composed of a disconnected union of smaller connected graphs. As shown in Figure 4, the second eigenvalue $\lambda_2 = 5.61$ is close to the predicted value $(n(p-q)/2 + (p+q)/(p-q) = 5.36)$, and can still be separated from the semi-circle bulk, and therefore the recovery is possible from v_2, the corresponding eigenvector shown in Fig. 5. We note that the largest eigenvalue, $\lambda_1 = 7.47$ is also very close to the predicted value, $n(p+q)/2 + 1 = 7.5$.

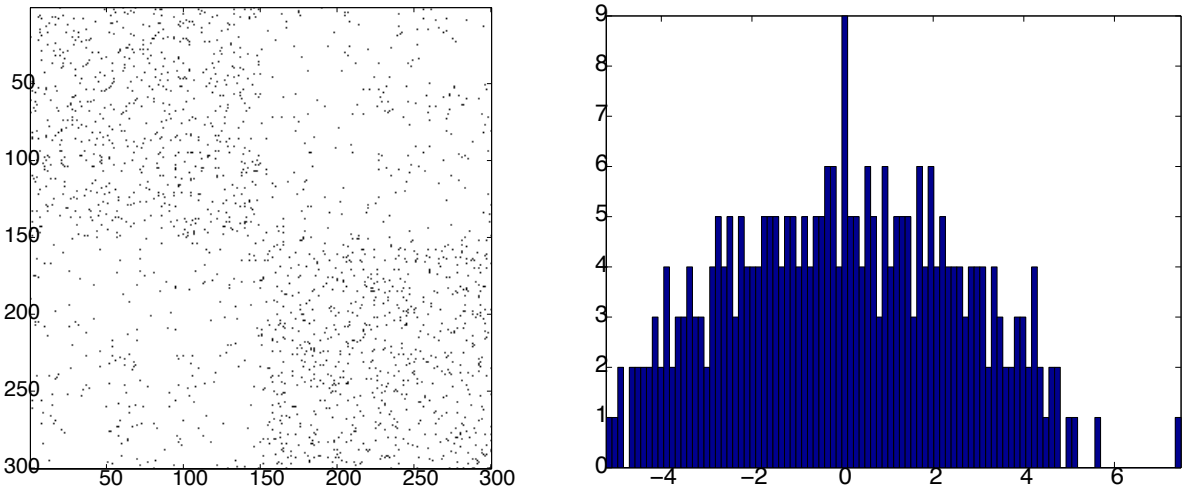

Figure 4: Left: Left: adjacency matrix A of a very sparse network; right: histogram of the eigenvalue of A. Note that the second largest eigenvalue can be separated from the bulk semi-circle.

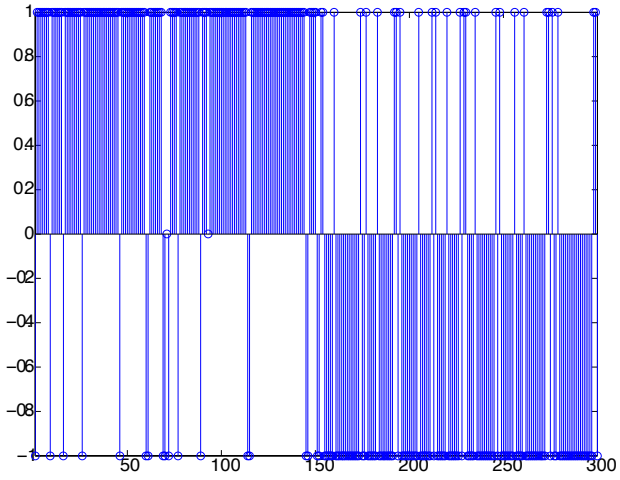

Figure 5: Eigenvector v_2 associated with the second largest eigenvalue λ_2 of the adjacency matrix shown in Fig. 4. Note that in this case `overlap = 0.7`.

17. For $n = 300$, and for each combination of $a = 5, 6, \ldots, 70$ and $b = 1, 2, \ldots, 50$ generate 20 random realizations of the planted partition model. Use the algorithm `Partition` to detect the communities, and compute the overlap score `overlap`.

    You will represent the results of this experiment by constructing a matrix `score`(α, β) that will contain the mean (computed over the 20 experiments) of the `overlap` score.

    Display the matrix in grayscale using the MATLAB function `imagesc`, and overlay the implicit curve defined by (22) that defines the condition on α on β for the detectability of the communities (see Fig. 6-right for an example of the figure).

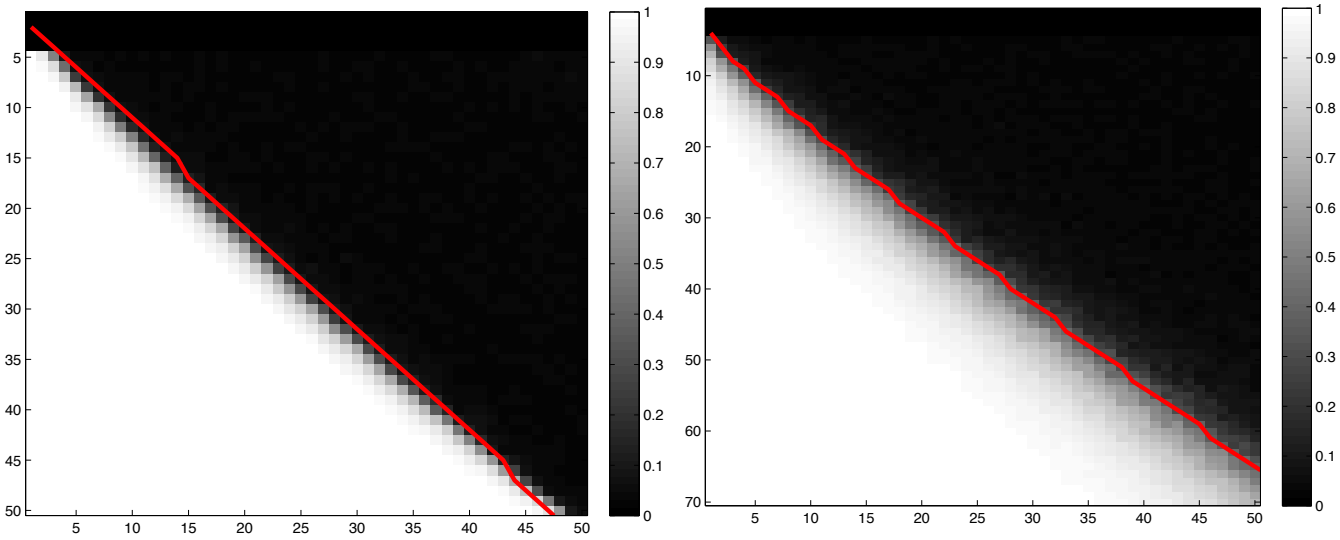

Figure 6: Probability of successfully detecting the partitions using the algorithm `Partition`. Left: dense network; right: sparse network. The x-axis corresponds to β (left) and b (right). The y-axis corresponds to α (left) and a (right). The decision boundaries defined by the equations (20) and (22) are overlaid in red. The recovery of the communities (measured by `overlap`) follows a phase transition (abruptly goes from 1 to 0) when the conditions are not met.

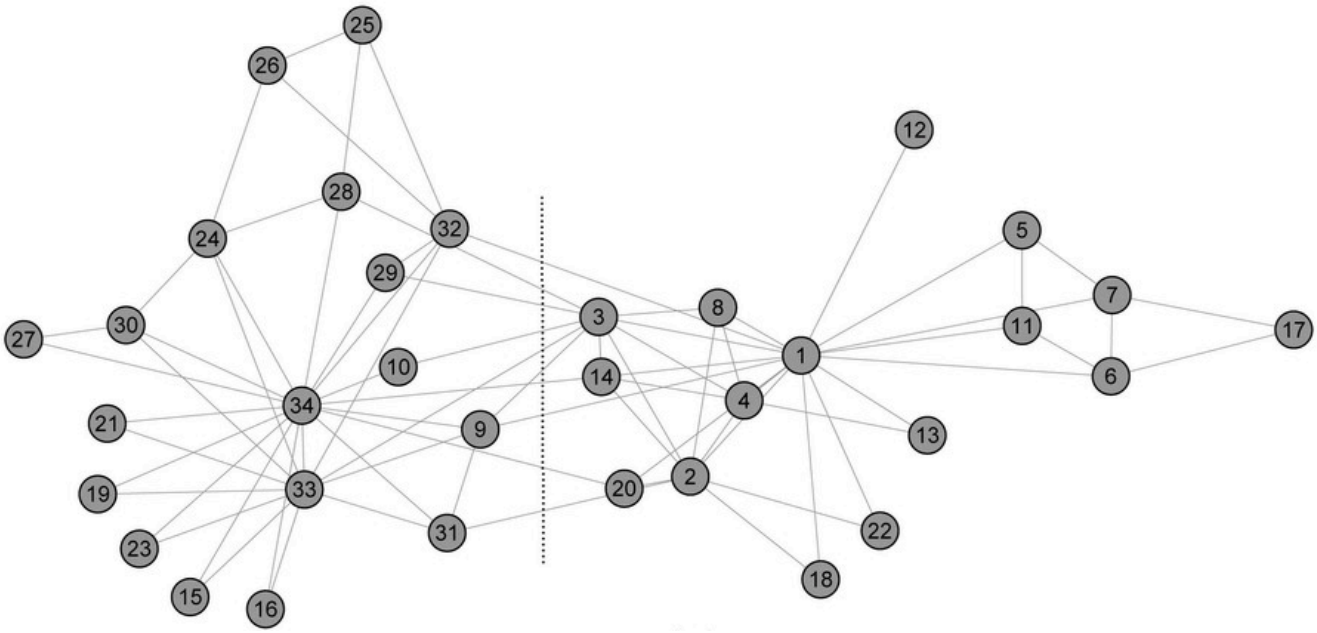

Figure 7: The Zachary's karate club network is split into two communities defined by the left and the fight of the vertical line.

An example of a social network

In this section, you will validate your algorithm on a real social network. This dataset is a standard dataset for all community detection algorithms. The Zachary's karate club network (see Fig. 7) is composed of 34 members of a karate club at a US university that was studied in the 1970. After a dispue between the manager (node 34) and the coach (node 1), the split into two groups, and the members joined the group in which they had the most friends. The cut between the two communities is illustrated by the dashed line in Fig. 7.

18. Load the matrix A from the file
    http://ece.colorado.edu/~fmeyer/class/ecen5322/zachary.mat
    and identify the two communities using the algorithm `Partition`.

19. Compute the overlap between the true partition ω, and the recovered partition $\widetilde{\omega}$.