Università della Svizzera italiana | Institute of Computing CI

**Numerical Computing** – Fall Semester 2024
**Lecturer:** Dr. Edoardo Vecchi
**Assistants:** Gianmarco De Vita, Samuele Pasini

# Project 2 – Social Networks

Due date: Wednesday, 16 October 2024, 11:59 PM

The purpose of this project[1] consists in the application of sparse linear algebra algorithms to the solution of problems pertaining to social network analysis.

## 1. Social Networks and the Householder XII Meeting

In June 1993, the UCLA Lake Arrowhead Conference Center in the San Bernardino Mountains, 90 miles east of Los Angeles, was the site of the Householder XII Symposium on Numerical Linear Algebra, organized by Gene Golub[2] and Tony Chan. Nick Trefethen posted a flip chart with Gene Golub's name in the center. He invited everyone present to add their name to the chart and draw lines connecting their name with the names of all their coauthors. The diagram grew denser throughout the week. At the end it was a graph with 104 vertices (or people) and 211 edges. John Gilbert entered the names and coauthor connections into Matlab, creating an adjacency matrix $A$. Using Matlab, we can start by retrieving the names and matrix stored in the PARC archive (`housegraph.mat`), which is available on iCorsi.

```
load housegraph
```

The variable `who` shows which variables are stored in the data file.

```
who
```

```
Your variables are:
A          Bunch          Funderlic  Ipsen       Moler       Reichel
ATrefethen BunseGerstner  George     Jessup      MuntheKaas  Ruhe
Ammar      Byers          Gilbert    Kagstrom    NHigham     Saied
Anjos      Chandrasekaran Gill       Kahan       NTrefethen  Sameh
Arioli     Crevelli       Golub      Kaufman     Nachtigal   Saunders
Ashby      Cullum         Gragg      Kenney      Nagy        Schreiber
Bai        Davis          Greenbaum  Kincaid     Ng          Smith
Barlow     Demmel         Gu         Kuo         Nichols     Starke
Benzi      Dubrulle       Gutknecht  Laub        OLeary      Stewart
Berry      Duff           Hammarling LeBorne     Ong         Strakos
Bjorck     Edelman        Hansen     Liu         Overton     Szyld
Bjorstad   Eisenstat      Harrod     Luk         Paige       TChan
Bojanczyk  Elden          He         Marek       Pan         Tang
Boley      Ernst          Heath      Mathias     Park        Tong
Boman      Fierro         Henrici    Meyer       Plemmons    VanDooren
Borges     Fischer        Hochbruck  Modersitzki Pothen      VanHuffel
VanLoan    Varah          Varga      Warner      Widlund     Wilkinson
Wold       Young          Zha        name        prcm        xy
```

and `name` shows the participants in the order they were added to the list:

```
name
```

```
name =
Golub
Wilkinson
TChan
```

[1] This assignment is originally based on a blog by Cleve B. Moler—who wrote a fantastic blog post about the Lake Arrowhead graph—and John Gilbert—who initially created the coauthor graph from the 1993 Householder Meeting. You can find more information here.

[2] `http://en.wikipedia.org/wiki/Gene_H._Golub`

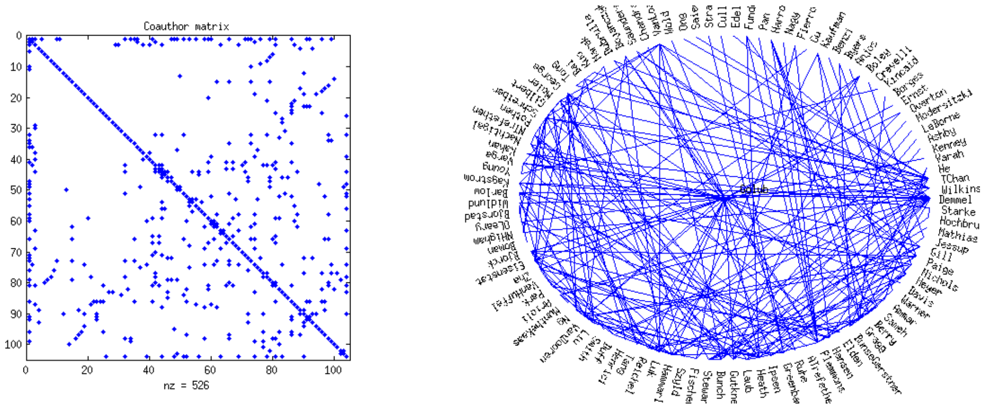

Figure 1: [Left] The coauthor matrix from the meeting (created with `spy(A)`, [Right] The coauthor circle from the meeting.

```
He
Varah
Kenney
Ashby
..
..
```

```
size(A)
```

```
ans =
104   104
```

Matrix $A$ is 104-by-104 and symmetric. Elements $A_{i,j}$ and $A_{j,i}$ are both equal to one if the people associated with indices $i$ and $j$ are coauthors and equal to zero otherwise. Matrix A is sparse, since most of the attendees are not coauthors and, thus, their corresponding entries have value 0. In order to check the matrix sparsity – i.e., the fraction of nonzero entries over the total number of elements – we can use the following command:

```
format short
sparsity = nnz(A)/prod(size(A))
```

```
sparsity =
0.0486
```

As we can observe, only roughly 5% of the entries of matrix A are different from 0 or, in other words, only 5% of all the possible pairs of attendees actually identify coauthors. In the left panel of Figure 1, we show a graphical representation of matrix A by means of a `spy` plot. The latter shows the location of the non-zeros, with Gene Golub in the first row/column, followed by the other authors in the order in which they appeared on Trefethen's flip chart.

## 1.1. Most prolific author

As the adjacency matrix is loaded from the archive, the most prolific coauthor is already in the first column. It was not a surprise to the conference participants that the most prolific coauthor is Gene Golub, one of the organizers.

```
m = find(sum(A) == max(sum(A)))
name(m,:)
```

```
m =
1
ans =
Golub
```

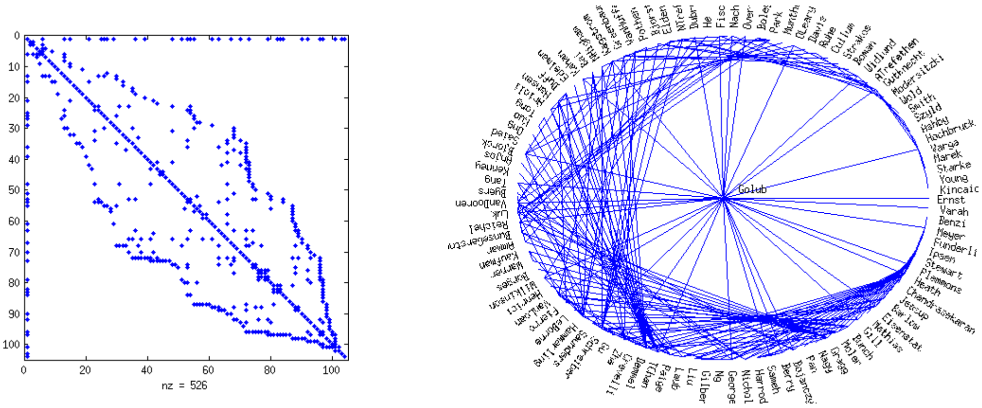

Figure 2: [Left] The reordered coauthor matrix $PAP^T$ from the meeting, [Right] The reordered coauthor circle.

### 1.2. Circular plot of all the authors

In the right panel of Figure 1, we created a circular plot with Golub in the center, the other authors around the circumference, and edges connecting the coauthors. If we place the authors around the circumference in the order we retrieve them from the chart, the edges would cross the circle pretty much randomly, as shown in Figure 1.

```
drawit;
snapnow
```

### 1.3. Reorder the authors

We want to rearrange the authors so that the coauthor connections are as close as possible to the circumference of the circle. This corresponds to a symmetric permutation of matrix $A$, aimed at minimizing, or at least reducing, its bandwidth. In 1969, Elizabeth Cuthill and John McKee described a heuristic for reordering the rows and columns of a matrix to reduce its bandwidth, now known as the *reverse Cuthill–McKee ordering*.

```
r = symrcm(A(2:end,2:end));
prcm = [1 r+1];
spy(A(prcm,prcm))
title('Coauthor matrix with reduced bandwidth');
drawitrcm;
snapnow
```

Figure 2 shows the impact of the reordering $P$ on both matrix $PAP^T$ and on the circular plot.

## 2. Spectral graph partitioning and the Laplacian with Matlab

In this segment, we will plant an artificial partition in a graph, and then use the eigenvector $v_2$ which is associated to the second smallest eigenvalue $\lambda_2$ to find it. The eigenvector $v_2$ or *Fiedler vector* plays an important role in graph partitioning. Its entries indicate a partitioning of the connectivity graph of the matrix it is associated with. In practice, this means that all indices corresponding to vector entries larger than zero belong to one set and all indices corresponding to vector entries smaller than zero belong to the other. The arising partition minimises the number of edges between the two sets. If you just keep reading on, we will walk you through the appropriate steps and you will see it for yourself in a practical application. If you would like to read more about this topic, a starting reference can be *Fiedlers theory of spectral graph partitioning*, by Slininger, available on iCorsi. We will start by generating a dataset. In this example, we will be a little more ambitious and use a larger number of vertices.

```
n=1000;
```

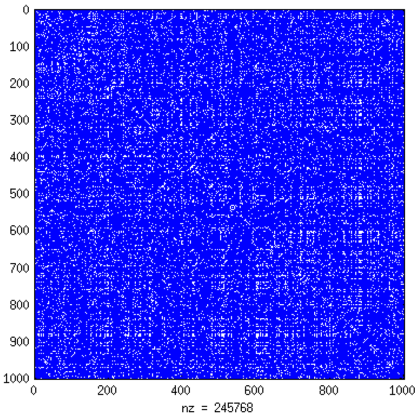

Figure 3: An almost dense artificial dataset.

To plant an artificial partition in our test dataset, we need to determine the vertices in each of the two groups. To do this, we randomly generate a permutation of all the vertices and select one set as group 1 and the rest as group 2. The variable gs controls how big the size of the first group is.

```
x = randperm(n);
gs = 450;
group1 = x(1:gs);
group2 = x(gs+1:end);
```

Now, we need to decide on the probabilities of edges within each group and between the two groups. Because we are planting a partition, the probabilities of edges between the groups should be much lower than the probability of edges within each group. Suppose that group 1 is a little more tightly connected than group 2. (Please insert your own amusing names for an actual identification of group 1 and group 2, e.g. politicians and mathematicians.)

```
p_group1 = 0.5;
p_group2 = 0.4;
p_between = 0.1;
```

With these probabilities in hand, we can construct our graph. The last few operations symmetrize the matrix.

```
A(group1, group1) = rand(gs,gs) < p_group1;
A(group2, group2) = rand(n-gs,n-gs) < p_group2;
A(group1, group2) = rand(gs, n-gs) < p_between;
```

Next, let's see if we can see the partition.

```
spy(A);
```

Figure 3 shows the result of the spy command. While some might still claim to see a partition in this data, we could argue that it is not particularly obvious. Now, let us investigate what the eigenvector $v_2$ tells us about this graph. We will use the eig command in Matlab to compute all eigenvectors and eigenvalues, and then store them as columns of matrix $V$ and as diagonal entries of matrix $D$ respectively.

```
A = triu(A,1);
A = A + A';
deg = sum(A);
L = diag(deg)-A;
[V, D] = eigs(L, 2, 'SA');
D(2,2)
ans =
199.9732
```

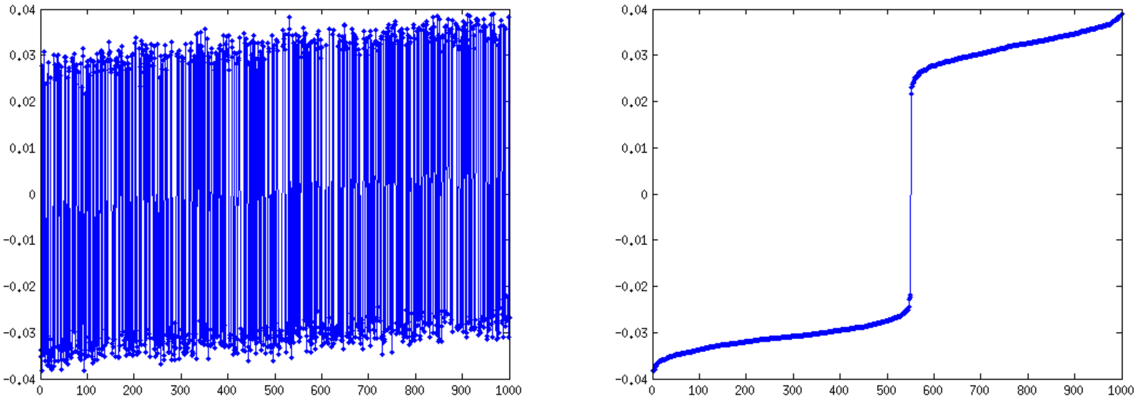

Figure 4: [left] The second smallest eigenvector, [right] The second smallest sorted eigenvector.

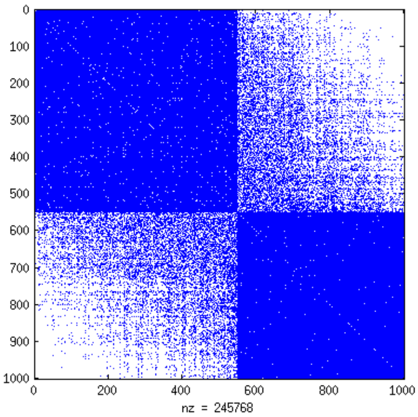

Figure 5: The results after the partitioning of our artificial dataset.

The second smallest eigenvalue $\lambda_2$ is greater than 0 if and only if we are considering a connected graph and its magnitude gives us an indication on how well-connected the overall graph is. To see what we found, let us plot the associated eigenvector $v_2$, corresponding to the second column $V(:,2)$ of the matrix defined above.

```
            plot(V(:,2), ’.-’);
```

The result is shown in the left part of Figure 4. The picture is not very useful in helping us understand how the graph is connected. However, we can try to sort the entries of $v_2$ and see if we can extract more information. The result is shown in the right part of Figure 4

```
            plot(sort(V(:,2)), ’.-’);
```

Now, this picture is much more informative than the previous one! We can notice a large gap in the middle of the sorted values. Interestingly enough, the number of points to the right of the gap is the same as `gs`, the size of our planted group. Let us see what happens when we permute the vertices of the graph according to this ordering.

```
            [ignore p] = sort(V(:,2));
            spy(A(p,p));
```

In Figure 5 we can finally clearly observe our partitioning. In the spectral analysis of a graph, the second smallest eigenvalue $\lambda_2$ is **very helpful** in finding a partitioning that splits the graph into **two subgraphs** while at the same time **minimizing the edges** between these two subgraphs.

Università della Svizzera italiana | Institute of Computing CI

**Numerical Computing** – Fall Semester 2024
**Lecturer:** Dr. Edoardo Vecchi
**Assistants:** Gianmarco De Vita, Samuele Pasini

## Solve the following problems [85 points]:

### 0. Preliminary: Read “A First Course on Numerical Methods”

Read section 5.7 on *Permutations and ordering strategies*, pp. 122–127 from the textbook [1]. Read the introductory section 1 (pp. 1–9), and section 3.3 (pp. 28–33) on spectral bisection from the report “Graph Partitioning: A survey” by Ulrich Elsner [2]. This last reference is useful alsoi for the next mini-project, where we will consider in more detail various graph partitioning techniques.

### 1. The reverse Cuthill–McKee ordering [10 points]

Load matrix ”A_SymPosDef.mat” from the dataset. You can reorder (permute) the matrix using the *reverse Cuthill–McKee ordering* via the Matlab function `symrcm()`, see Matlab documentation for more information. Visualize both the original and reverse Cuthill–McKee permuted matrix and comment on what you observe. Compute the Cholesky factor of the original matrix and the permuted matrix. Visualize the Cholesky factors and comment on the number of nonzeros.

### 2. Sparse matrix factorization [20 points]

Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, with entries $A_{ij}$ defined as follows:

$$A_{ij} = \begin{cases} 1, & \text{if } i = 1 \text{ or } i = n \text{ or } j = 1 \text{ or } j = n \text{ and } i \neq j \\ n + i - 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Please note that the increasingly larger values on the diagonal are necessary to ensure the positive definiteness of matrix A. Answer the following questions:

1. Construct matrix $A$ for the case $n = 10$ and explicitly write down its entries. How many non-zero elements does it have?

2. We now want to derive a general formula to compute the number of non-zero entries. Show that, for a given matrix $A \in \mathbb{R}^{n\times n}$ with this structure, the number of non-zero elements is $5n - 6$.

3. Write a function `A_construct()`, which takes as input $n$ and returns, as output, the matrix $A$ defined in Eq. 1 and its number of non-zero elements *nz*. Test your function in a script `ex2c.m` for $n = 10$ and compare your results with those you obtained in point (a). Furthermore, within the same script, visualise the non-zero structure of matrix $A$ by using the command `spy()`.

4. Using again the `spy()` command, visualize side by side the original matrix $A$ and the result of the Cholesky factorization (`chol()` in Matlab). Comment on the results obtained.

5. Explain why, for $n = 100,000$, using `chol()` to solve $Ax = b$ for a given right-hand-side vector $b$ would be problematic. Are there ways to mitigate this issue?

### 3. Degree centrality [5 points]

In graph theory and network analysis, centrality refers to indicators which identify the most important vertices within a graph. Applications include identifying the most influential person(s) in a social network, key infrastructure nodes in the Internet or urban networks, and super spreaders of disease. Here we are interested in the **Degree centrality**, which is conceptually simple. It is defined as the number of links incident upon a node (i.e., the number of vertices

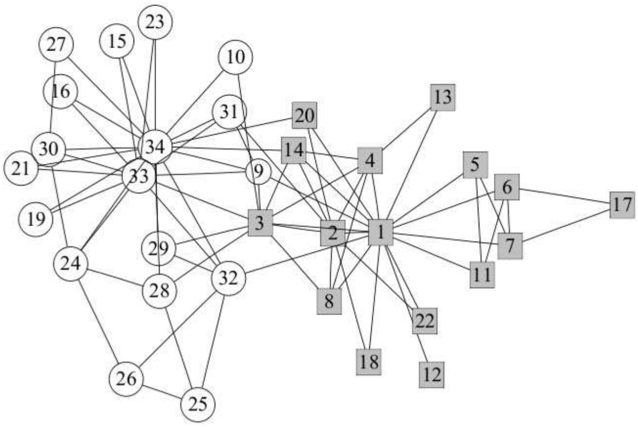

Figure 6: The social network of a karate club at a US university.

that a node has). The degree centrality of a vertex $v$, for a given graph $G := (V, E)$ with $|V|$ vertices and $|E|$ edges, is defined as the numbers of edges of vertex $v$. Compute the degree centralities for the top 5 authors. Include them in an ordered list, and show the authors, their coauthors and the degree centrality.

## 4. The connectivity of the coauthors [10 points]

How many coauthors do the authors have in common? Think about a general procedure that allows you to compute the list of common coauthors of two authors and express it in matrix notation (carefully explain your reasoning in the report). Use the formula you derived to compute the common coauthors of the pairs (Golub, Moler), (Golub, Saunders), and (TChan, Demmel). Who are these common coauthors? Report their names.

## 5. PageRank of the coauthor graph [5 points]

Compute the PageRank value (e.g., by using a modified version of `pagerank.m` from Project 1) for all authors and provide a graph of all authors in descending order according to the PageRank. Include your script in the submission.

## 6. Zachary's karate club: an introduction to spectral graph partitioning [35 points]

Figure 6 shows the social network of a karate club at a US university in the 1970s[3]. At the end of the study, a conflict between club members arose. As a consequence, the club formally split into two separate organizations (white circles vs. grey squares). Please use the adjacency matrix `karate.adj` and answer the following numerical questions.

1. Write a Matlab code that ranks the five nodes with the largest degree centrality? What are their degrees?
2. Rank the five nodes with the largest **eigenvector centrality**. What are their (properly normalized) eigenvector centralities?[4]
3. Are the rankings in (a) and (b) identical? Were you expecting different results? Carefully describe the similarities and the differences and comment the results obtained.
4. Use spectral graph partitioning to find a near-optimal split of the network into two groups of 16 and 18 nodes, respectively. List the nodes in the two groups. How does spectral bisection compare to the real split observed by Zachary (see Fig. 6)? Comment on the results obtained.

[3] Image from: M. E. J. Newman and M. Girvan, Phys. Rev. E 69, 026113 (2004).

[4] Eigenvector centrality is defined by the principle introduced in the previous assignment. The eigenvector returned by the PageRank algorithm corresponds to the weights associated with each node.

## Quality of the code and of the report [15 Points]

The highest possible score for each project is 85 points and up to 15 additional points can be awarded based on the quality of your report and code (maximum possible grade: 100 points). Your report should be a coherent document, structured according to the template provided on iCorsi. If there are theoretical questions, provide a complete and detailed answer. All figures must have a caption and must be of sufficient quality (include them either as .eps or .pdf). If you made a particular choice in your implementation that might be out of the ordinary, clearly explain it in the report. The code you submit must be self-contained and executable, and must include the set-up for all the results that you obtained and listed in your report. It has to be readable and, if particularly complicated, well-commented.

## Additional notes and submission details

Summarize your results and experiments for all exercises by writing an extended LaTeX report, by using the template provided on iCorsi. Submit your gzipped archive file (tar, zip, etc.) **on iCorsi strictly before the deadline** in the dedicated section and use the following standard naming: `project_1_lastname_firstname.zip` (or `tgz`). Submission by email or through any other channel will not be considered. Late submissions will not be graded and will result in a score of 0 points for that project. You are allowed to discuss all questions with anyone you like, but: (i) your submission must list anyone you discussed problems with and (ii) you must write up your submission independently. Please remember that plagiarism will result in a harsh penalization (0 points) for all involved parties and that the usage of generative AI, even for rephrasing, is strictly forbidden.

## In-class assistance

If you experience difficulties in solving the problems above, you can receive in-class assistance either on Tuesdays (13:30-15:00, in room D1.14) or on Wednesdays (13:30-15:00, in room D0.03). Please refer to this schedule for any eventual change in the allocated room.

## References

[1] Ascher, U. M., & Greif, C. (Eds.). (2011). *A first course on numerical methods*. SIAM.

[2] Elsner, U. (1997). *Graph partitioning-a survey*.