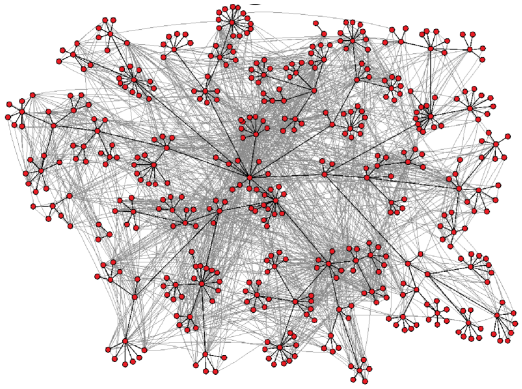
# Social Network Data Analysis

MIE223

Winter 2025

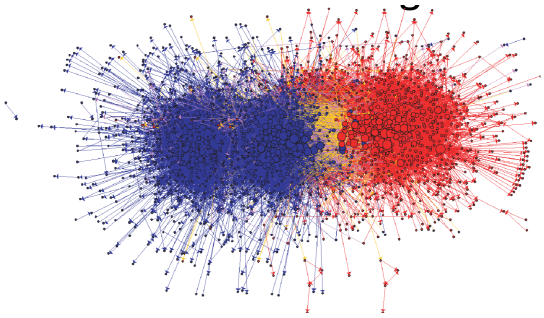
## 1 Social Network Analysis

### 1.1 Networks: An organization - HP Labs



not a lot of inter-team communication as they are present together. the network represents the email communication between the teams. holistic interconnectivity exists where the center dot is. there is a clear tree hierarchy, with clumpiness where certain nodes have high degree connections. social networks are not random, they are structured.

### 1.2 Political blogs



the network represents the links between political blogs. the network is divided into two main groups, with a few blogs connecting the two groups. there are a lot of echo chambers.

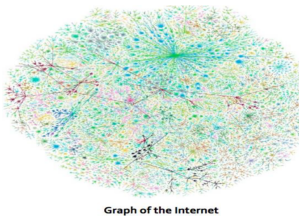
### 1.3 A View of Facebook via 10 M links

data is generated by a process, which is subject to selection biases.

### 1.4 Why Networks?

Behind each of these complex systems there is an intricate wiring diagram, a network that defines the interactions between system components.
Understanding the network is key to understand the behaviors of such complex system.

### 1.5 Networks




Communications


Transportation

### 1.6 Application domains in network analysis

- Social (people-people) networks (social network)
- Information networks (social network)
- Organization and political networks (social network)
- Computer networking
- Biology
- Transportation networks

## 2 EXAMPLES OF PROBLEMS IN SOCIAL NETWORKS

### 2.1 E.g. 1: Small-world phenomena

how do you decide which nodes connect to what in a social network? there is a tree pattern in the connections, where there are supernodes with multiple connections and other nodes where they only have a few connections. you are more likely to friend a node if that node already has a lot of friends.
everyone is at most 60 degrees of connection away from each other
there is a power law connection in how nodes with many connections tend to gain more connections. these have a natural hierarchal structure and are not random.

### 2.2 E.g. 1 (cont.): Link prediction

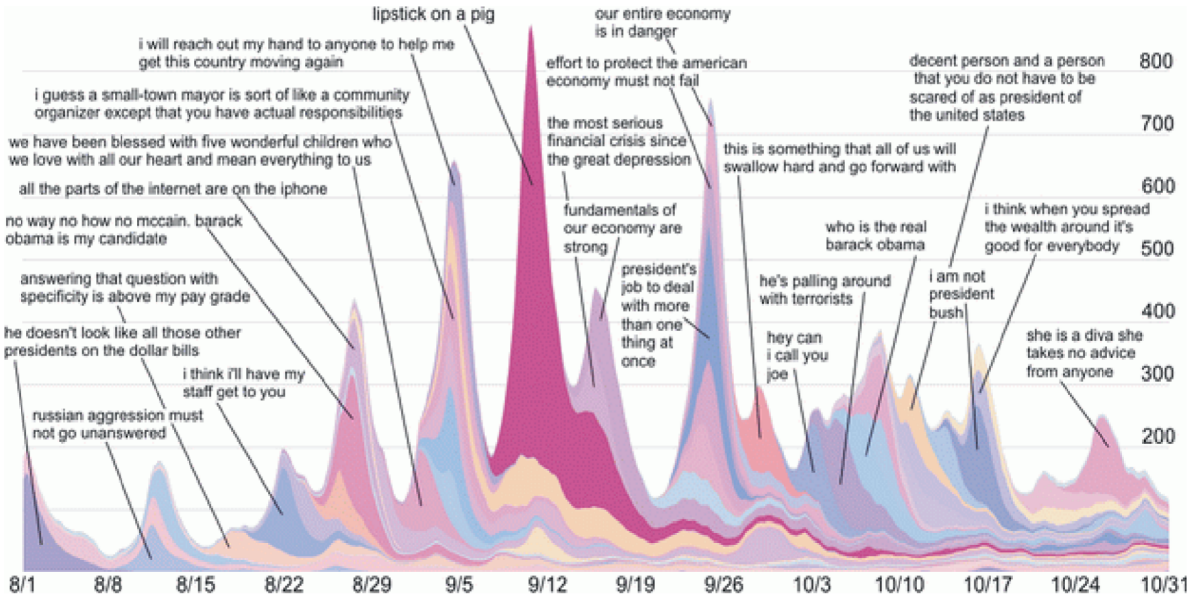
“Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future?”

- What are the measurements?
- – “proximity” and “similarity” between two unconnected nodes

- What are application domains?
- – social networks, friend recommendation; product / webpage recommendation; predicting academic collaborations; predicting merger and acquisitions ...
- How to measure performance?
- – use future held-out data, conversion rate, ..

there is a connectivity such that you are more likely to connect to people your friends are connected to

## 2.3 E.g. 2: Tracking Memes – What goes Viral?




what is it about ”lipstick on a pig” that makes it go viral? it is novel, it is surprising, it is interesting, it is complex, it is emotional it is a story, it is a narrative, it is a meme.

## 2.4 E.g. 3: Community detection

- Why detect communities?
- Simplified network structure and “big picture”
- Explain actions and positions
- Better predictions
- What defines a community?
- Interaction
- Profile
- Dynamics
- ...

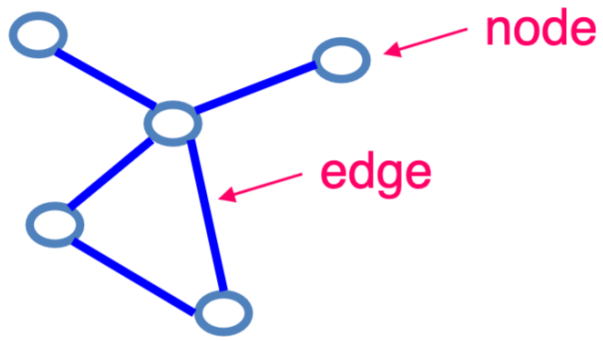
# 3 Social Network Analysis: Preliminaries

## 3.1 What are networks?

**Note 1.** Networks are sets of nodes connected by edges.

"Network" = "Graph"

| points | lines | |
|---|---|---|
| vertices | edges, arcs | math |
| nodes | links | computer science |
| sites | bonds | physics |
| actors | ties, relations | sociology |




## 3.2 Network elements: edges

Directed (also called arcs, links): A-¿B
Undirected: A¡-¿B or A - B
Undirected is a subset of directed such that all undirected links have directed links

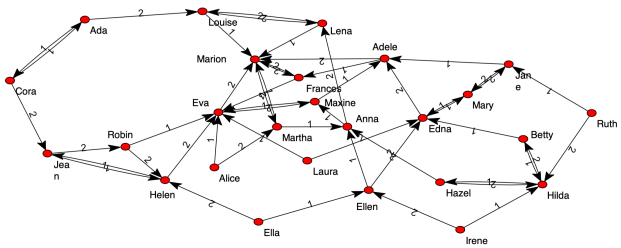
## 3.3 Edge attributes

Examples

- weight (e.g. frequency of communication)
- ranking (best friend, second best friend...)
- type (friend, relative, co-worker)
- properties depending on the structure of the rest of the graph: e.g. betweenness

## 3.4 Directed networks

Girls' school dormitory dining-table partners, 1st and 2nd choices (Moreno, The sociometry reader, 1960)




### 3.5 Document Elements In Twitter

- User
- Mention
- Hashtag
- Hyperlink

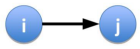
## 4 Data representation

1. Adjacency matrix
2. Edge list
3. Adjacency list

### 4.1 (1) Adjacency matrices

Representing edges (who is adjacent to whom) as a matrix

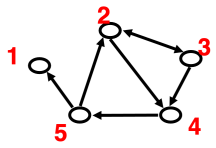
$A_{ij}$ = 1 if node *i* has an edge to node *j*
= 0 if node *i* does not have an edge to *j*




$A_{ii}$ = 0 unless the network has self-loops


$A_{ij} = A_{ji}$ if the network is undirected,
or if *i* and *j* share a reciprocated edge


### 4.2 (1) Adjacency matrices example




A =

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |

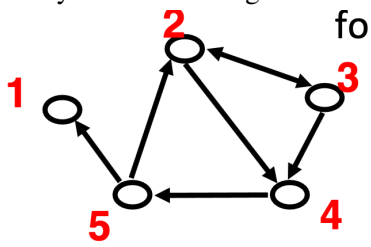
### 4.3 (2) Edge list

- 2, 3
- 2, 4
- 3, 2

- 3, 4
- 4, 5
- 5, 2
- 5, 1

### 4.4 (3) Adjacency lists

are easier to work with if network is • large • sparse
quickly retrieve all neighbors for a node




for a node

- 1:
- 2: 3 4
- 3: 2 4
- 4: 5
- 5: 1 2

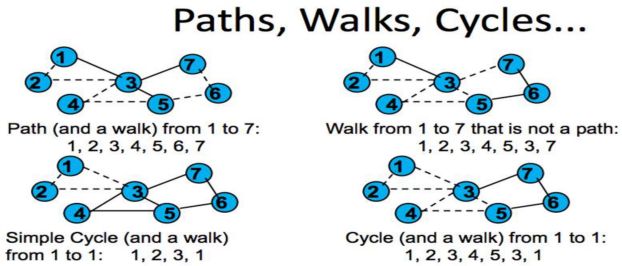
stored as a key (dictionary) instead of a list. use a sparse matrix in python to store this.

## 5 Computing metrics

1. Path and distance
2. In/Out degree
3. Centrality

### 5.1 Distances in a Network

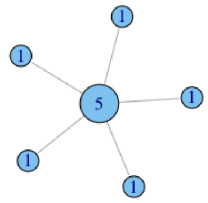
- Path: a walk (i1,i2,... ik) with each node ij distinct
- Cycle: a walk where i1 = ik
- Geodesic: a shortest path between two nodes




## 6 Who is the Center of a network?

### 6.1 Local view of the Social Network

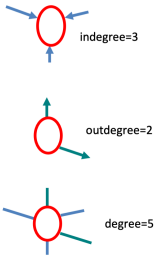
**Note 2.** One centrality definition: Nodes with more friends are more central.




Assumption: the connections that your friend has don't matter, it is what they can do directly that does (e.g. go have a beer with you, help you build a deck...)

## 6.2 Indegree and Outdegree

- Indegree
  - how many directed edges (arcs) are incident on a node
- Outdegree
  - how many directed edges (arcs) originate at a node
- Degree (in or out)
  - number of edges incident on a node




## 6.3 Node degree from matrix values

- Outdegree = $\sum_{j=1}^{n} A_{ij}$

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row* $\sum_{j=1}^{n} A_{3j}$

- Indegree = $\sum_{i=1}^{n} A_{ij}$

example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column* $\sum_{i=1}^{n} A_{i3}$

A =

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |

A =

| | | | | |
|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |