Graph Mining
SD212
1. Graphs as sparse matrices

Thomas Bonald

2019 – 2020

Motivation

Real graphs are **sparse**

Dataset	#nodes	#edges	Density
Flights	2,939	30,500	$\approx 10^{-3}$
Amazon products	335k	925k	$\approx 10^{-5}$
Actors	382k	33M	$\approx 10^{-4}$
Wikipedia (en)	12M	378M	$\approx 10^{-6}$
Twitter	42M	1.5G	$\approx 10^{-6}$
Friendster	68M	2.5G	$\approx 10^{-7}$

Outline

1. Sparse matrices
2. Graphs as sparse matrices
3. The friendship paradox

Sparse matrices

$$\begin{bmatrix} 5 & 6 & 9 & 0 & 2 & 2 & 0 & 4 \\ 7 & 0 & 0 & 0 & 7 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 5 & 5 \\ 5 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 5 & 0 & 0 & 0 & 9 & 0 \end{bmatrix}$$

Coordinate format

$$\begin{bmatrix} 5 & 6 & 9 & & 2 & 2 & & 4 \\ 7 & & & & 7 & & & \\ & & 5 & & & & 5 & 5 \\ 5 & & & & & 3 & & \\ 6 & & & & & & & 3 \\ & & 5 & & & & 9 & \end{bmatrix}$$

$$\begin{aligned} \text{data} &= (5, 6, 9, 2, 2, 4, 7, 7, 5, 5, 5, 5, 3, 6, 3, 5, 9) \\ \text{row} &= (0, 0, 0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 3, 4, 4, 5, 5) \\ \text{col} &= (0, 1, 2, 4, 5, 7, 0, 4, 2, 6, 7, 0, 5, 0, 7, 2, 6) \end{aligned}$$

Compressed Sparse Row

$$\begin{bmatrix} 5 & 6 & 9 & & 2 & 2 & & 4 \\ 7 & & & & 7 & & & \\ & & 5 & & & & 5 & 5 \\ 5 & & & & & 3 & & \\ 6 & & & & & & & 3 \\ & & 5 & & & & 9 & \end{bmatrix}$$

$$\begin{aligned} \text{data} &= (5, 6, 9, 2, 2, 4, 7, 7, 5, 5, 5, 5, 3, 6, 3, 5, 9) \\ \text{indices} &= (0, 1, 2, 4, 5, 7, 0, 4, 2, 6, 7, 0, 5, 0, 7, 2, 6) \\ \text{indptr} &= (0, 6, 8, 11, 13, 15, 17) \end{aligned}$$

Compressed Sparse Column

$$\begin{bmatrix} 5 & 6 & 9 & & 2 & 2 & & 4 \\ 7 & & & & 7 & & & \\ & & 5 & & & & 5 & 5 \\ 5 & & & & & 3 & & \\ 6 & & & & & & & 3 \\ & & 5 & & & & 9 & \end{bmatrix}$$

$$\begin{aligned} \text{data} &= (5, 7, 5, 6, 6, 9, 5, 5, 2, 7, 2, 3, 5, 9, 4, 5, 3) \\ \text{indices} &= (0, 1, 3, 4, 0, 0, 2, 5, 0, 1, 0, 3, 2, 5, 0, 2, 4) \\ \text{indptr} &= (0, 4, 5, 8, 8, 10, 12, 14, 17) \end{aligned}$$

List of Lists

$$\begin{bmatrix} 5 & 6 & 9 & & 2 & 2 & & 4 \\ 7 & & & & 7 & & & \\ & & 5 & & & & 5 & 5 \\ 5 & & & & & 3 & & \\ 6 & & & & & & & 3 \\ & & 5 & & & & 9 & \end{bmatrix}$$

$$\text{data} = [[5, 6, 9, 2, 2, 4], [7, 7], [5, 5, 5], [5, 3], [6, 3], [5, 9]]$$
$$\text{rows} = [[0, 1, 2, 4, 5, 7], [0, 4], [2, 6, 7], [0, 5], [0, 7], [2, 6]]$$

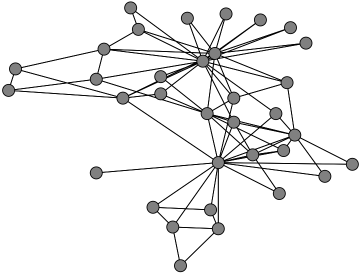
Use cases

Fast...	COO	CSR	CSC	LIL
Dot product		✓	✓	
Arithmetic		✓	✓	
Row slicing		✓		
Column slicing			✓	
Modification				✓
Loading	✓			

Outline

1. Sparse matrices
2. **Graphs as sparse matrices**
3. The friendship paradox

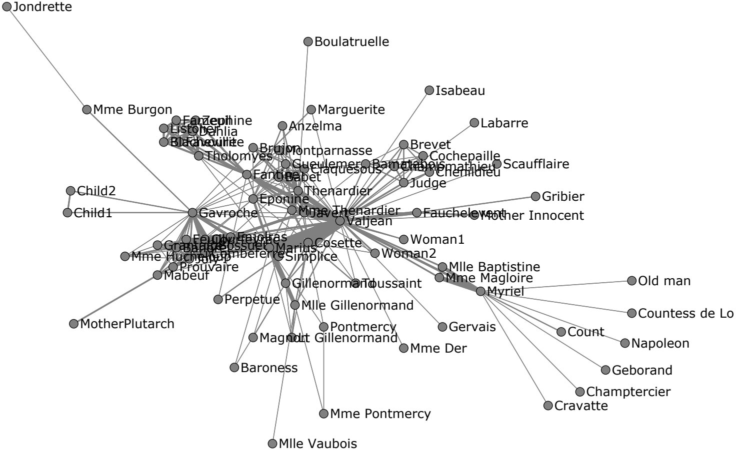
Graph



Adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

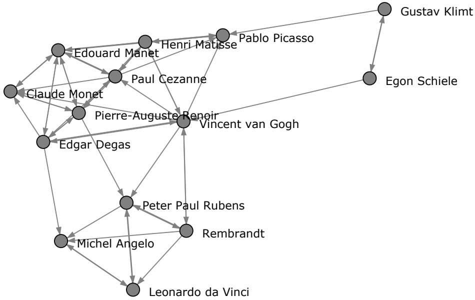
Weighted graph



Adjacency matrix:

$$A_{ij} = \begin{cases} w_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

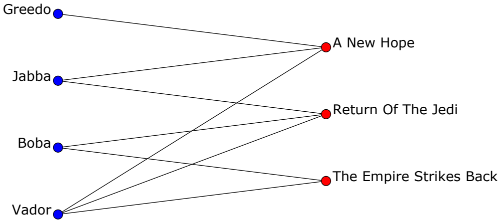
Directed graph



Adjacency matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$$

Bipartite graph



Biadjacency matrix:

$$B_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

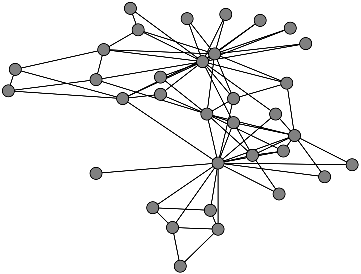
Outline

1. Sparse matrices
2. Graphs as sparse matrices
3. **The friendship paradox**

Your friends have more friends than you on average.

Node sampling

Consider a graph of n nodes and m edges, without self-loops.

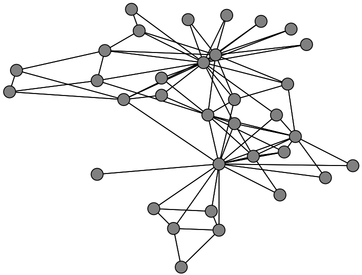


Let X be a random node and D its degree.

$$\mathrm{E}_0(D) = \frac{2m}{n}$$

Edge sampling

Select one of the edges, uniformly at random:



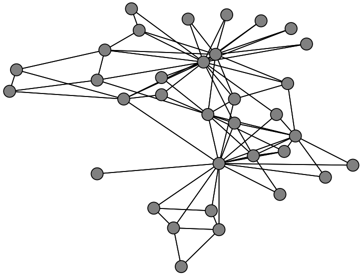
Bias

The degree D has the size-biased distribution.
In particular,

$$\mathrm{E}_\infty(D) \geq \mathrm{E}_0(D)$$

Neighbor sampling

Select a node, then one of its neighbors uniformly at random:

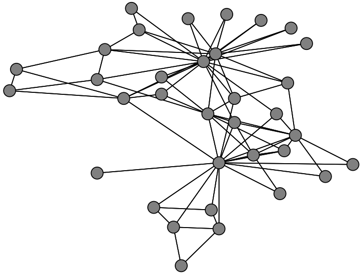


The friendship paradox

$$\mathrm{E}_1(D) \geq \mathrm{E}_0(D)$$

Random walk

Select a node, then walk t steps at random:



Let π_t be the distribution of X, as a row vector:

Stationary distribution

If the graph is connected and not bipartite,

$$\lim_{t \to +\infty} \pi_t = \frac{d^T}{2m}$$