MU5IN075
Network Analysis and Mining
1. Introduction

Esteban Bautista-Ruiz, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

first_name.last_name@lip6.fr

September 14, 2021

How does it work?

Teachers

Esteban Bautista-Ruiz esteban.bautista-ruiz@lip6.fr
Lionel Tabourier lionel.tabourier@lip6.fr

Webpage

http://lioneltabourier.fr/teaching_en.html

Language

Course in English, questions in both English and French

How does it work? (2)

Prerequisites

- Programming (in python, jupyter environment)
- Notions in graph algorithmics is a plus

Topics

From theoretical tools to practical problems:

- Introduction to complex networks analysis (LT, 4 weeks)
- Graph models (LT, 2 weeks)
- Communities (EBR, 2 weeks)
- Learning on graphs (EBR, 2 weeks)
- Web algorithms (LT, 3 weeks?)
- Structure of the Internet (LT, 1 week?)

How does it work? (3)

Calendar

See:
https://cal.ufr-info-p6.jussieu.fr/master/
http://planning.upmc.fr/jussieu/M2.INFO.RES/

Careful:

- 14 Sep: Course only
- 21 Sep to 16 Nov: Lab then Course
- 30 Nov to 11 Jan: Course then Lab
- 18 Jan: Lab only

Location:

- Course changes (check online)
- Lab in 14-15/509

How does it work? (4)

Assessment

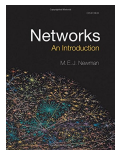
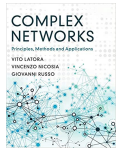
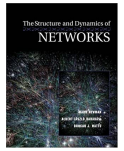
- Lab works (every week): 50%
- Exam: 50% (Careful: content of assignments $\neq$ exam!)

About Lab works

- around 12 Lab works to send back
- by groups of 1 or 2 (but no more)
- 5 of them (randomly drawn) are marked
- you are allowed 2 jokers ($\Rightarrow$ you can send back 10)

About textbooks

No definitive textbooks, but a few recommendations (week 1–8)

Networks, An Introduction.	M.Newman	
Complex Networks.	V.Latora V.Nicosia G.Russo	
The Structure and Dynamics of Networks.	Research papers selection	

Outline

Outline

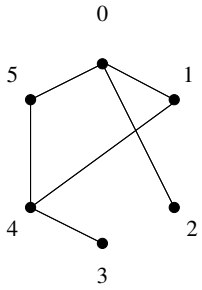
Graph

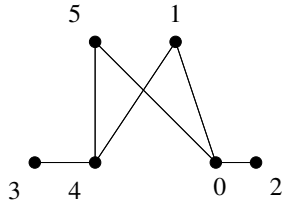
A graph $G = (V, E)$ is a couple of sets.

- V is a set of *vertices* (or *nodes*) ***fr:*** *sommet, nœud*
- $E \subseteq (V \times V)$ is a set of *edges* (or *links*) ***fr:*** *lien, arête*

Example

- $V = \{0, 1, 2, 3, 4, 5\}$
- $E = \{(0, 1), (0, 2), (3, 4), (4, 5), (5, 0), (1, 4)\}$





Warning: The graph should not be confused with its drawing!

Graph theory

Graph theory is an important and well studied field:
`https://en.wikipedia.org/wiki/Graph_theory`

When was it founded?
circa 1736, Euler's 7 bridges of Königsberg problem.

Why is this useful to us?
Many systems can be modeled using graphs and we can use graph theory to study them!

Graph theory

Graph theory is an important and well studied field:
`https://en.wikipedia.org/wiki/Graph_theory`

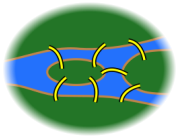
When was it founded?
circa 1736, Euler's 7 bridges of Königsberg problem.

Why is this useful to us?
Many systems can be modeled using graphs and we can use graph theory to study them!

Graph theory

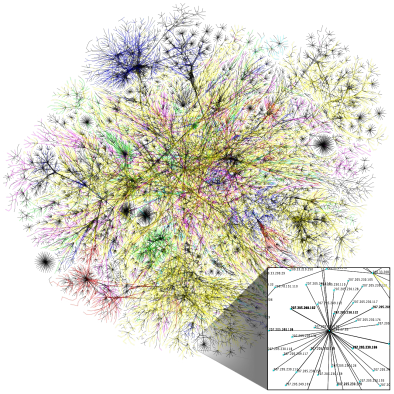
Graph theory is an important and well studied field:
`https://en.wikipedia.org/wiki/Graph_theory`

When was it founded?
circa 1736, Euler's 7 bridges of Königsberg problem.



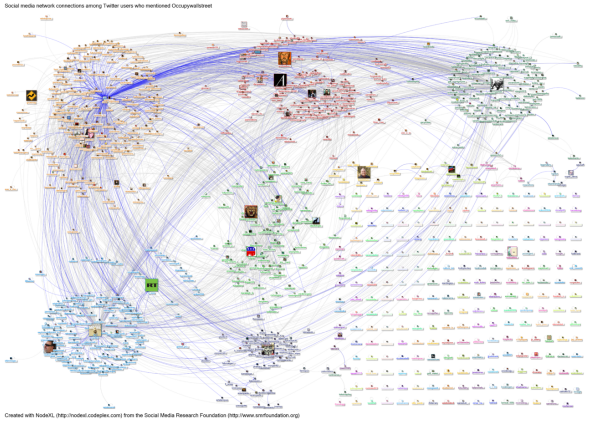
Why is this useful to us?
Many systems can be modeled using graphs and we can use graph theory to study them!

Internet: computers connected by internet connections

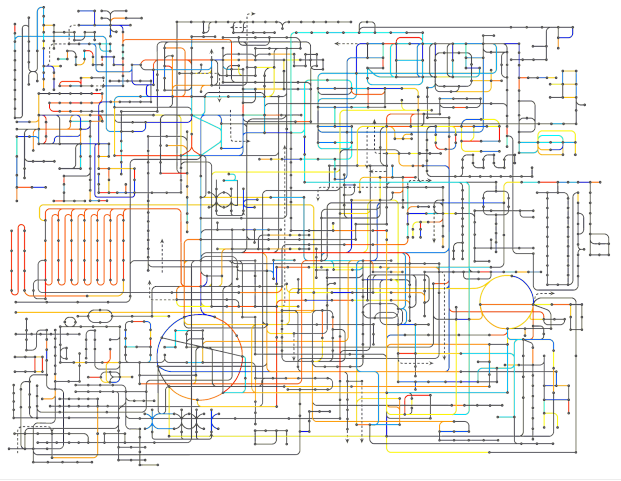


At the IP (or at the AS) level

Twitter: profiles connected by mentions and replies



Metabolism: proteins interacting with each other



More examples

field	network	node	edge
Telecom	Internet (1) Internet (2)	router autonomous system	IP network adjacency BGP connection
Information	WWW Wikipedia document network	web page Wikipedia page article, patent	hyperlink hyperlink citation
Social/Eco	OSN trade disease	account entity person	friendship good exchange contaminate
Biology	metabolic brain	protein neuron	chemical reaction synaptic connection
Transport	air transport railway network road network power grid	airport station road relay station	direct connection railroad intersection power line
Linguistics	co-occurrence	word	same sentence

These graphs extracted from the "real-world" are called complex networks or real-world graphs.

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

What is the point for a Master RES student?

Objective (assessment)

- master basic tools of complex network analysis both theoretical (course) and practical (TP)
- analyze research results in CNA independently

But really...

- know more about graph representation of networks
- ...and basic python programming on the topic
- some basics of data mining (of course not comprehensive)
- general culture in computer science
- train a critical eye on a problem from theory to practice

Outline

Often very large

- Facebook has more than 2 billion active accounts.
- Several billion routers on the Internet
- A human brain has 10^{11} (100 billion) neurons and more than 10^{14} (100 trillion) synapses.

We note:

- $n = |V|$ the number of nodes
- $m = |E|$ the number of edges

Often sparse

Not many edges compared to number of edges that could exist.

Maximum number of edges of an undirected graph of n nodes:

$$m_{max} = ??$$

Ratio between the number of edges and the maximum number of edges of an undirected graph of n nodes and m edges:

$$\delta = ??$$

= probability that a random link exists

Often sparse

Not many edges compared to number of edges that could exist.

Maximum number of edges of an undirected graph of n nodes:

$$m_{max} = \frac{n \cdot (n-1)}{2}$$

Ratio between the number of edges and the maximum number of edges of an undirected graph of n nodes and m edges:

$$\delta = \frac{2 \cdot m}{n \cdot (n-1)}$$

= probability that a random link exists

Connectedness

Path *(chemin)* from u to v: sequence of edges $(u, v_1), (v_1, v_2), \ldots, (v_{k-1}, v)$

Length *(longueur)* = k (number of edges in a path)

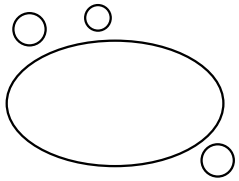
Connected component *(composante connexe)*: maximal set of nodes such that $\exists$ a path between all pairs of nodes.

Connected graph: only one connected component

Connectedness

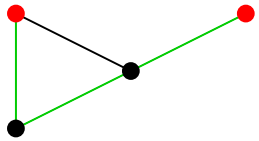
For complex networks

In general, a giant component
$\rightarrow$ contains most nodes



Distance

Path from u to v = sequence of edges u...v

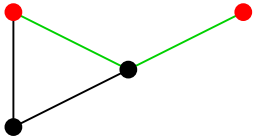


a path of length 3

Distance

Path from u to v = sequence of edges u...v

distance $d(u, v)$ = length of *one* shortest path



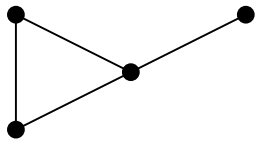
shortest path of length 2 $\Rightarrow$ distance = 2

Distance

Path from u to v = sequence of edges u...v

distance $d(u, v)$ = length of *one* shortest path

diameter Δ = longest distance between all pairs of nodes



diameter = 2

Distances and connectedness

Average distance: average distance for all pairs of nodes

Connectedness?
Distance defined for two nodes of the same connected component

In practice

Average distance, diameter:
→ in the largest connected component

Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

Distances and connectedness

Average distance: average distance for all pairs of nodes

Connectedness?
Distance defined for two nodes of the same connected component

In practice

Average distance, diameter:
→ in the largest connected component

Distance

For complex networks

In general, short distances ($\sim \log(n)$)

example: actor collaborations, Kevin Bacon game
`https://oracleofbacon.org/`

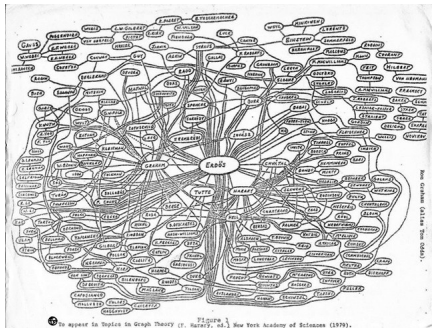
Graphs and networks
Common properties of networks
Applications
Graph algorithmics: storage in memory

Distance

For complex networks

In general, short distances ($\sim \log(n)$)

scientific collaborations, Erdős number:



Milgram experiments

The first small-world experiments

The small-world problem, *Psychology today*, 1967

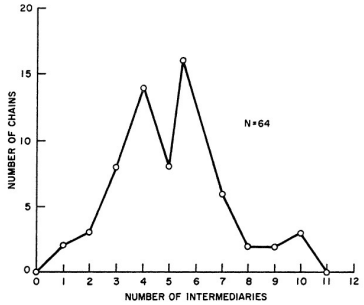
Two studies, one protocol:

- several sources (Wichita, Kansas / Omaha, Nebraska)
- one target (Cambridge / Sharon, Massachusetts)
- information: target name, occupation (housewife / broker)
- **rule: pass on a file to a unique acquaintance**

experimenter follows who sent the file to whom

A first picture of small-worlds

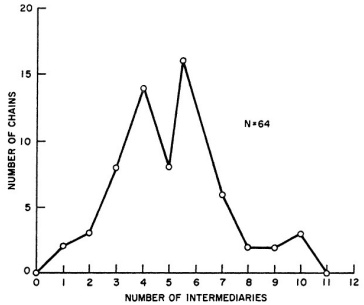
Most famous result: median = 5 intermediaries
(counter-intuitive?)



Quasi-geometric increase of the number of relations
⇒ *"Six degrees of separation"* (not Milgram's words)

A first picture of small-worlds

Most famous result: median = 5 intermediaries
(counter-intuitive?)



Quasi-geometric increase of the number of relations
⇒ *"Six degrees of separation"* (not Milgram's words)

Degree

u and v are neighbours if there is an edge between them.

undirected graph

Degree: $d(v)$ number of neighbours of v ***(fr:** degré)*

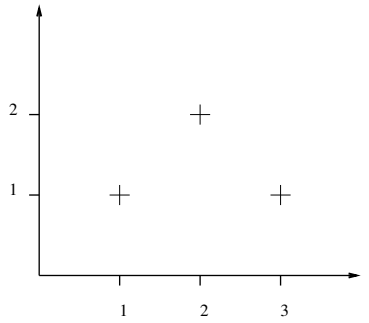
Degree distribution (1/2)

Degree distribution:
4 nodes, degrees: 2 2 3 1

Degree distribution (1/2)

Degree distribution:
4 nodes, degrees: 2 2 3 1
Distribution: how many nodes have degree k, as function of k.

$1 \to 1, 2 \to 2, 3 \to 1$



Power-law

Power-law *(loi de puissance)*

- $N_k \sim k^{-\alpha}$
- straight line in log-log scale

Heterogeneous distribution: close to a power-law (in our cases)

Power-law

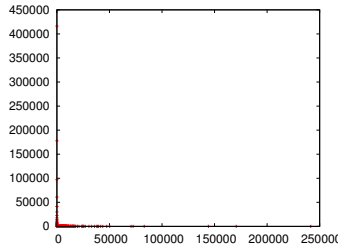
- line in log-log scale on several orders of magnitude

$\neq$

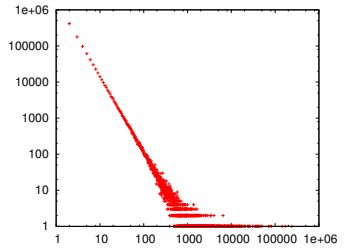
Heterogenous

- on several order of magnitude
- close to linear in log-log scale

Heterogeneous distributions: log-log scale

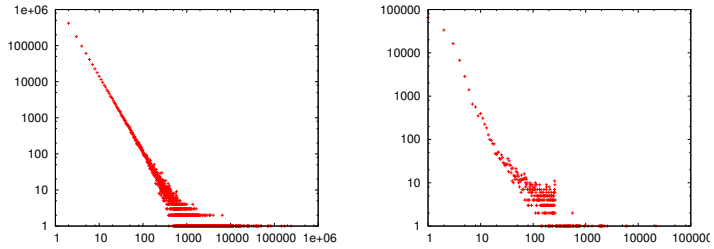


linear scale

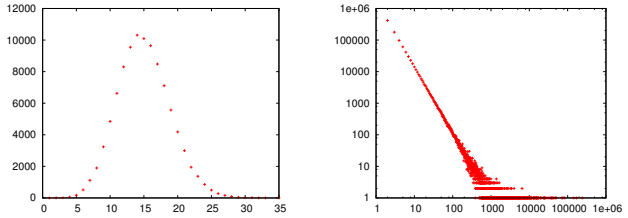


logarithmic scale

Examples



Heterogeneous vs homogeneous distributions



Homogeneous

Idea of normality (and exceptions)

Heterogeneous

Any kind of behaviour exists → no simple idea of normality

Degree distributions (2/2)

For complex networks

In general, degree distributions are heterogeneous;
there exist hubs (nodes with many connections).

Clustering coefficient

clustering coefficient $\mathrm{cc}(v)$

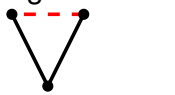
= probability that two neighbours of v are connected

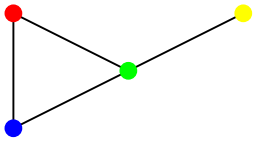
= # pairs of connected neighbours / # pairs of neighbours
= local density

Clustering coefficient

clustering coefficient cc(v)

= probability that two neighbours of v are connected





clustering coefficient: 1, 1, $\frac{1}{3}$, undefined

Clustering coefficient

Clustering coefficient of a graph:
average on all the nodes of degree ≥ 2

For complex networks

In general, high clustering

meaning several orders of magnitudes greater than density

Intuitively: the friends of my friends are my friends.

Transitive ratio

Another coefficient to measure the local density of a graph:
transitive ratio

$$\mathrm{tr}(G) = \frac{3N_\Delta}{N_\vee}$$

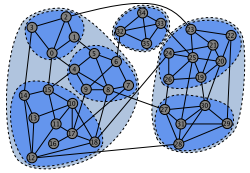
N_Δ: number of triangles of the graph
$N_\vee$: number of connected triplets

Careful: definitions vary in the literature
(sometimes called *global clustering*)

Defining communities

Goal

Looking for an internal structure of the graph.



Definition

- **intuitive:** people sharing a common interest, web pages having similar contents...
- **structural:** zone of the graph with a high density of links

Common properties – conclusion

Most complex networks share common properties:

size	large
density	low
connectivity	giant component
distances	low
degree distribution	heterogeneous
clustering	high
communities	yes

Let's check it:
`http://konect.cc/networks/`

Outline

Applications: some examples

- Represent data, locate information
  - Map a network of relationships
  - Use the map to find information, search engine
- Community detection
  - Organization of documents or friends' lists.
  - "People you may know" (or "You may also like").
- Routing
  - Making better routing protocols.
  - Improving transport systems.
- Spreading processes
  - Detecting influential spreaders.
  - Controlling the propagation of diseases.

Outline

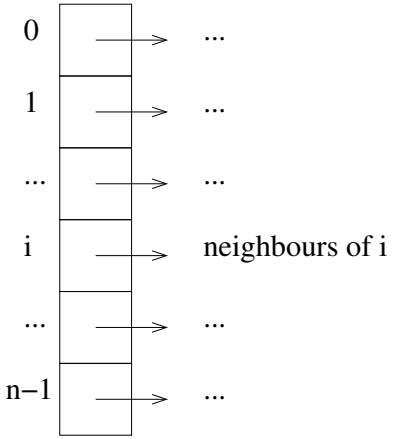
Adjacency matrix

$$\begin{array}{c|c} & \begin{matrix} 0 & 1 & \dots & n-1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ \dots \\ n-1 \end{matrix} & \begin{pmatrix} & & & \\ & & \dots & \\ & & & \end{pmatrix} \end{array}$$

Cell (i, j):

- 1 if there is an edge between i and j
- 0 otherwise

Adjacency list



Advantages, drawbacks

Temporal and spatial complexities

	Matrix	List
Existence of an edge		
List the neighbours of a node		
Size		

In general, real world networks have low density (sparse)
$\rightarrow m << n^2$
(we often consider that m is "a few times n")

Advantages, drawbacks

Temporal and spatial complexities

	Matrix	List
Existence of an edge	$\mathcal{O}(1)$	$\mathcal{O}(d(v))$
List the neighbours of a node	$\mathcal{O}(n)$	$\mathcal{O}(d(v))$
Size	$\mathcal{O}(n^2)$	$\mathcal{O}(m)$

In general, real world networks have low density (sparse)
$\rightarrow m << n^2$
(we often consider that m is "a few times n")