ELU 501
Data science, graph theory and social network studies

Yannis Haralambous (IMT Atlantique)

April 16, 2018

ELU 501
Data science, graph theory and social network studies

Yannis Haralambous (IMT Atlantique)

April 16, 2018

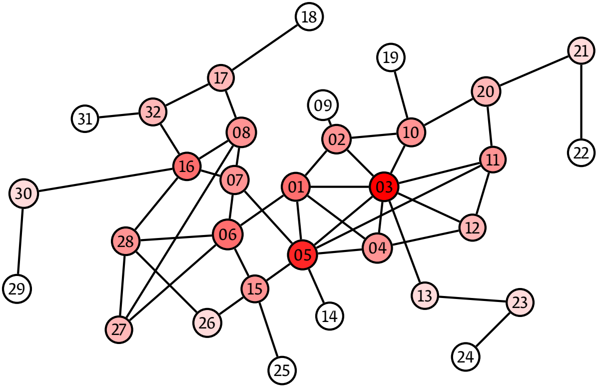
Part II

Lecture 2
Measures

Centrality

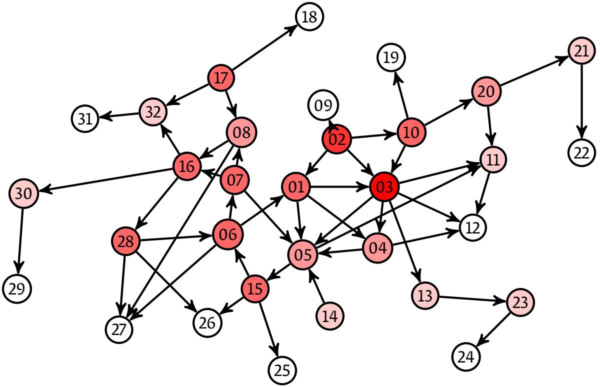
- In the following we will study graphs according to their metric properties.
- The first question we may ask is "which are the most important / most central vertices in a graph?"

Degree centrality



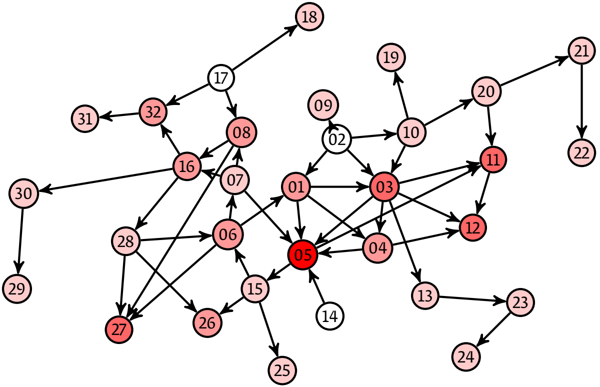
- Simplistic approach: *the most central vertex is the one of highest degree*: we call this measure, *degree centrality* ⟨*centralité par le degré*⟩.
- We calculate for each v_i the quantity $\frac{d(v_i)-d_{\min}}{d_{\max}-d_{\min}}$.
- In the figure, v_{14} has a very weak measure although it is a neighbor or v_5.

Outdegree centrality



- In the directed case we can consider *as central the vertex transmitting to a maximum of neighbors*: this is *outdegree centrality* ⟨*centralité par le degré sortant*⟩.
- We calculate for each v_i the quantity $\frac{d^+(v_i)-d^+_{\min}}{d^+_{\max}-d^+_{\min}}$.
- In the figure, v_{12} has a very weak measure although it is a neighbor or v_3.

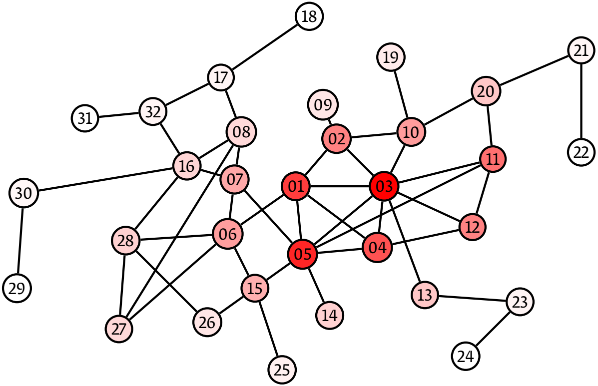
In-degree centrality



- We can also consider *as central the vertex which receives from a maximum of neighbors*: this is *indegree centrality* ⟨*centralité par le degré entrant*⟩.
- We calculate for each v_i the quantity $\frac{d^-(v_i)-d^-_{\mathrm{min}}}{d^-_{\mathrm{max}}-d^-_{\mathrm{min}}}$.
- In the figure, v_{14} has a very weak measure although it is a neighbor or v_5.

Eigenvector centrality

Yannis Haralambous (IMT Atlantique)



- More sophisticated: *consider as central the vertex having the most <u>central</u> neighbors*: we call it *eigenvector centrality* ⟨🇫🇷 *centralité par vecteur propre*⟩.
- Notice that v_{14} is a bit stronger than v_{25}, itself stronger than v_{22}.

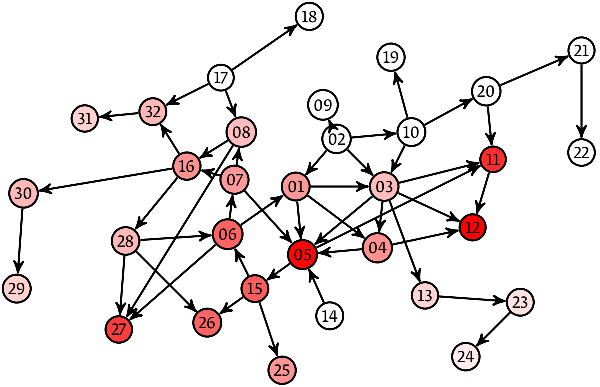
Eigenvector centrality

- But how do we calculate this kind of centrality, where we want the centrality of neighbors to be recursively taken into account?
  1. we attach value 1 to each vertex $x_i = 1$,
  2. we do a first calculation $x'_i = \sum_j a_{i,j} x_j$, i.e., $X' = A \cdot X$,
  3. after t steps we will have $X(t) = A^t \cdot X(0)$,
  4. like any vector, $X(0)$ can be written as a linear sum of eigenvectors of A: $X(0) = \sum_i c_i \vec{v_i}$,
  5. then, if κ_i is the ith eigenvalue (by decreasing order)

     $$X(t) = A^t \cdot \sum_i c_i \vec{v_i} = \sum_i c_i \kappa_i^t \vec{v_i} = \kappa_1^t \sum_i c_i \left(\frac{\kappa_i}{\kappa_1}\right)^t \vec{v_i},$$

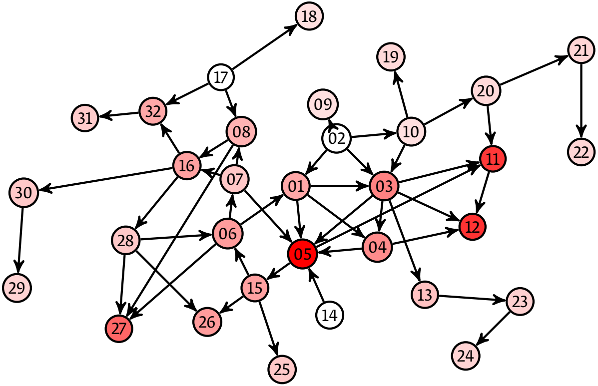
     (the second equality comes from the fact that if κ_i is an eigenvalue and v_i the corresponding eigenvector, then $(A - \kappa_i \mathrm{Id}) \cdot v_i = 0$ and hence $A v_i = \kappa_i v_i$ and $A^t v_i = \kappa_i^t v_i$),
  6. when $t \gg 0$, $X(t) \sim c_1 \kappa_1^t \vec{v_1}$ and hence centrality is proportional to the first eigenvector.
- We define eigenvector centrality of X as: $A \cdot X = \kappa_1 X$.

Entering eigenvector centrality



- We can generalize eigenvector centrality to directed graphs, but sometimes this is problematic.
- Here v_2 has zero entering centrality and therefore this also the case of v_{10}, etc., although these vertices have entering arrows...

Katz centrality

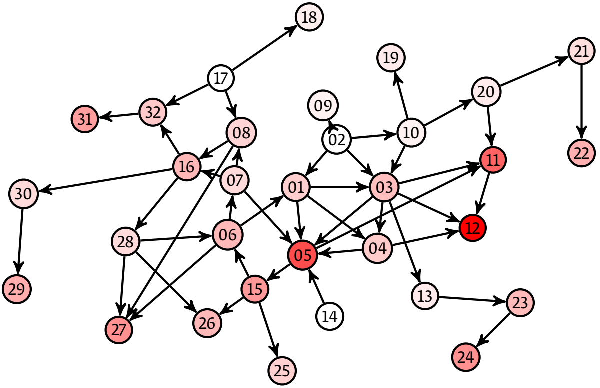


- Katz (1953) has improved entering eigenvector centrality, as we can see here only vertices without entering arrow have zero centrality.

Katz centrality

- Katz's idea: give a small amount of centrality to everyone and see how it evolves: $x'_i = \alpha \sum_j a_{i,j} x_j + \beta$.
- By setting $\beta = 1$ we get $X = (\mathrm{Id} - \alpha A)^{-1} \cdot 1$, *Katz centrality* ⟨*centralité de Katz*⟩.
- The choice of α is arbitrary, but we don't want the measure to diverge.
- It diverges when $\det(\mathrm{Id} - \alpha A) = 0$, i.e., $\alpha^{-1} = \kappa_*$, that is, for the first time for $\alpha = 1/\kappa_1$. So we set α slightly less than $1/\kappa_1$.
- In the example we took $\alpha = \frac{1}{4}$.
- To do calculations we will not invert A (complexity $O(n^3)$), but we will iterate $X' \leftarrow \alpha A \cdot X + \beta \cdot 1$.
- There is still a problem with Katz: if a vertex points to a million others, they will all get the same amount of centrality, while it would be better to disolve this centrality in the mass of entering arrows.

PageRank



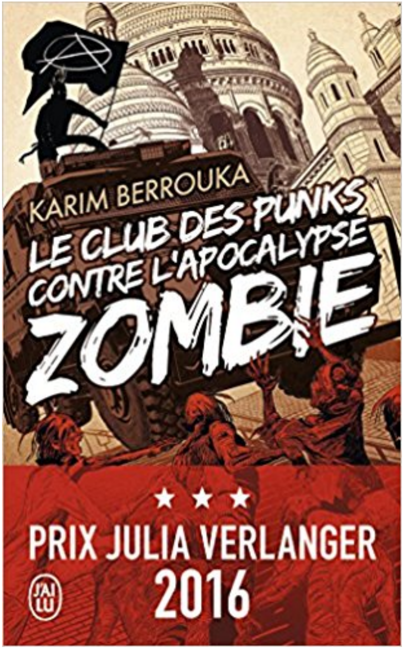
- A solution to this problem is given by the *PageRank* algorithm thanks to which Google is a major company today.
- Note that, contrarily to previous methods, isolated strings like $v_{13}v_{23}v_{24}$ ou $v_{30}v_{29}$) increase centrality. Note also that although v_3 and v_{12} both have 3 entering arrows, their centrality is quite different.

PageRank

- PageRank defines $x'_i = \alpha \sum_j a_{i,j} \frac{x_j}{d^+(v_j)} + \beta$, where $d^+(v_j)$ is the leaving degree of v_j.
- What happens when $d^+(v_j) = 0$? We replace it by $\max(d^+(v_j), 1)$.
- Let D be the diagonal matrix of the $\max(d^+(v_j), 1)$. Then we have $X = \alpha A \cdot D^{-1} \cdot X + \beta \cdot 1$ and hence for $\beta = 1$:
$$X = D \cdot (D - \alpha A)^{-1} \cdot 1.$$
- It is commonly agreed that the success of Google is not due to the relevance of the global set of results returned, but to their <u>order</u>.
- And behing the order you have PageRank.
- The same calculations as for Katz show that α must be less than 1. Google uses $\alpha = 0.85$.

PageRank and random walks

- Another interpretation of PageRank is centrality can be obtained by *random walks* ⟨🇫🇷 *marche aléatoire*⟩ :
  A random walk is a random path on a graph where at each moment t the probability of visiting any neighboring vertex is $1/d$ where d is the degree of the current vertex.
- To random walk we add another operation called *teleport* (*Energy Mr Sulu!*): on each vertex, there is
  1. an α probability that the next move will be some other (not necessarily neighboring) vertex, uniformily chosen among all vertices of the graph,
  2. and a $1-\alpha$ probability that the standard random walk is pursued.
- Compare this with zombies (or tourists) randomly walking through Paris.

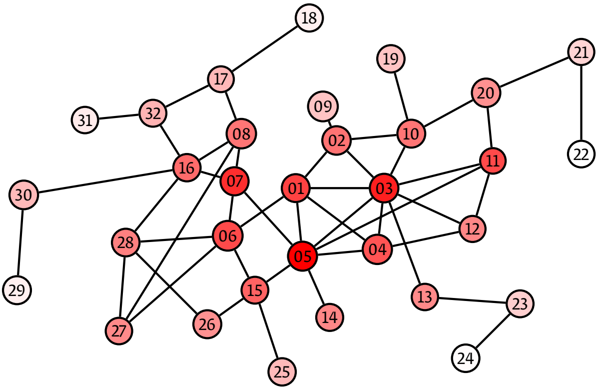

KARIM BERROUKA
LE CLUB DES PUNKS CONTRE L'APOCALYPSE
ZOMBIE
★ ★ ★
PRIX JULIA VERLANGER
2016
J'AI LU

Closeness centrality

- Let us return to more geometric considerations.
- We can calculate the *mean geodesic distance* ⟨*distance géodésique moyenne*⟩ of a vertex v_i out of all other vertices: $\ell_i = \frac{1}{n}\sum_j d(v_i, v_j)$.
- More we are central, more this quantity diminuishes.
- We define *closeness centrality* ⟨*centralité de proximité*⟩ by

$$C_i = \frac{1}{\ell_i} = \frac{n}{\sum_j d(v_i, v_j)}.$$

Closeness centrality



Closeness centrality

- We have calculated closeness centrality for actors of database `imdb.com` (where edges mean: actors have played in the same movie) and the most central acror is Christopher Lee, with $C = 0,4143$ and the less central actor is Leia Zanganeh with $C = 0,1154$.
- There is almost half a million actors between those two and nevertheless the two values are quite close.
- Additional problem: in an unconnected graph, all C are zero...
- And if we limit ourselves to components, we get greater values for small components.

Closeness centrality

- The solution to this problem is to use an harmonic mean:
$$C'_i = \frac{1}{n-1}\sum_{j\neq i}\frac{1}{d(v_i, v_j)}.$$
- The good properties are: if v_i and v_j belong to distinct components, the fraction is zero.
- Close to each other vertices count more than far away ones.
- We also define the *mean geodesic distance* ⟨🇫🇷 *distance géodésique moyenne*⟩
$$\ell = \frac{1}{n}\sum_i \ell_i = \frac{1}{n^2}\sum_{i,j} d(v_i, v_j)$$
- and the *harmonic mean geodesic distance* ⟨🇫🇷 *distance géodésique harmonique moyenne*⟩
$$\ell' = \frac{n}{\sum_i C'_i} = n(n-1)\frac{1}{\sum_{i\neq j}\frac{1}{d(v_i,v_j)}}.$$

Betweenness centrality

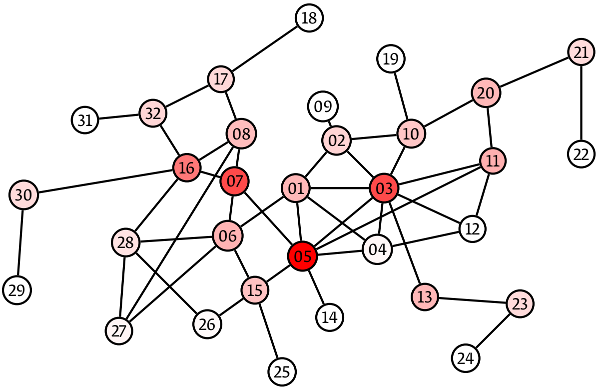
- In a network where travel ideas, opinions, merchandises, IP packets between close to everyone may not be the most relevant centrality characteristic.
- Idea: let us consider paths going through a vertex.
- We want to quantify the geodesic paths traversing a given vertex.
- we define *betweenness centrality* ⟨🇫🇷 *centralité de synexité*⟩ as $x_i := \sum_{j,k} \frac{n_{i,j,k}}{g_{j,k}}$ where:
  - $n_{i,j,k}$ is the number of geodesic paths between v_j and v_k going through v_i,
  - $g_{j,k}$ is the number of geodesic paths between v_j et v_k.

Betweenness centrality

Yannis Haralambous (IMT Atlantique)

- Betweenness centrality is completely different than other centralities since it is totally independent of degree.
- Example: if $G = G_1 - v_i - G_2$ where G_1 and G_2 are very populated components, then v_i will have a high betweenness centrality even though its degree is only 2.
- Such a vertex is called a *broker* ⟨*courtier*⟩.

Betweenness centrality



Betweenness centrality

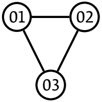
Yannis Haralambous (IMT Atlantique)

- Betweenness centrality values are distributed as follows:
- Maximal value is n^2 (think of the center of a star graph).
- Minimal value is $2n-1$ (think of the leaf of a linear graph, counting also a zero-length path).
- In IMDB, the actor of greatest betweenness centrality is Fernando Rey, he played both with US and European actors, in movies and TV, with $x = 7.47 \cdot 10^8$. The second one is again Christopher Lee with $x = 6.46 \cdot 10^8$, a difference of 14%.
- BC is much more stable than CC.
- We can also define *normalized betweenness centrality* ⟨*centralité de synexité normalisée*⟩, with values between 0 and 1:

$$x_i = \frac{1}{n^2} \sum_{j,k} \frac{n_{i,j,k}}{g_{j,k}}.$$

Transitivity

- In algebra, a relation $*$ is transitive when $a * b, b * c \Rightarrow a * c$.
- In social networks this can be interpreted as "the friend of my friend is my friend".
- A complete subgraph K_3 is called a *triad* ⟨*triade*⟩.



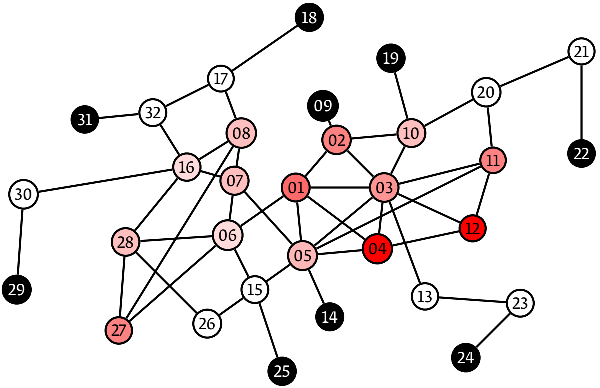
- We define the *global clustering coefficient* ⟨*coefficient de clustering global*⟩ of a graph as the ratio

$$C = \frac{\#\text{triads}}{\#\text{paths of length } 2}.$$

- For a given vertex v_i of degree ≥ 2, we define the *local clustering coefficient* ⟨*coefficient de clustering local*⟩ :

$$C_i = \frac{\#\text{connected neighbors of } v_i}{\#\text{pairs of neighbors of } v_i}.$$

Transitivity



Black vertices are of degree 1 and hence C_i cannot be defined. Note that v_4 and v_{12} are the "most transitive" vertices of the graph.