PMI RESEARCH & DEVELOPMENT

Théorie des graphes et biologie moléculaire : la sociologie des gènes et des protéines

Florian Martin

Maths et Société, Université de Neuchâtel
25 novembre 2015

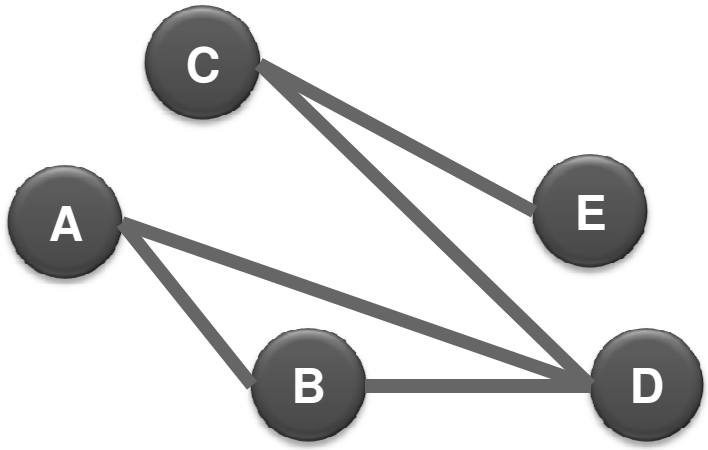
Agenda

- Théorie des Graphes et Réseaux
- Dogme Central de la Biologie Moléculaire et Biologie des Systèmes
- Hubs et Réseaux invariants d'échelle ("scale-free")
  - Les réseaux biologiques sont-ils aléatoires?
- Communautés dans les réseaux
  - Une méthode spectrale
- Frustration: "*Les amis de mes amis sont mes amis et les ennemis de mes amis sont mes ennemis*".

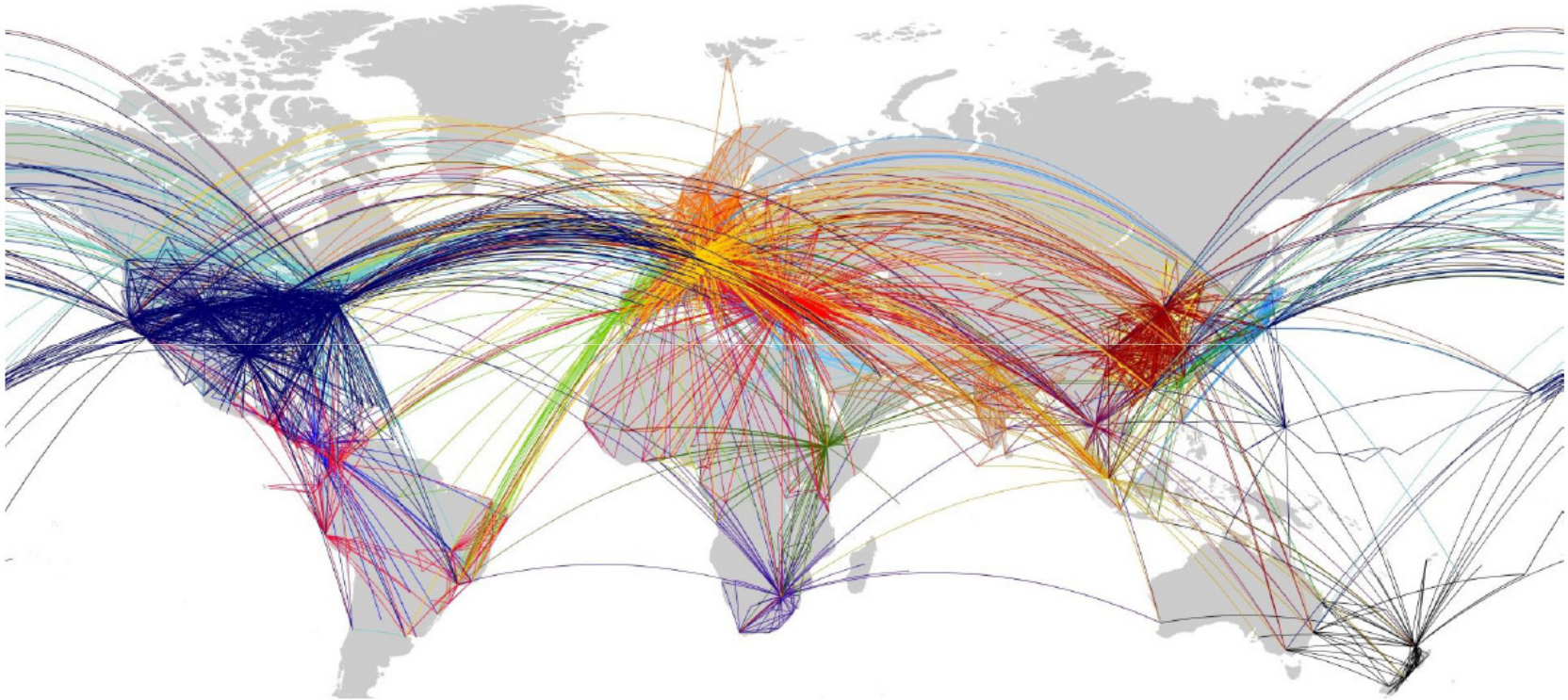
Définition d'un graphe

- Graphe: Un graphe G est un ensemble de sommets V, et un ensemble d'arêtes $E \subset V \times V$ reliant chacun deux sommets.
- Les arêtes peuvent être dirigées, pondérées et parfois même signées.

- Exemple:



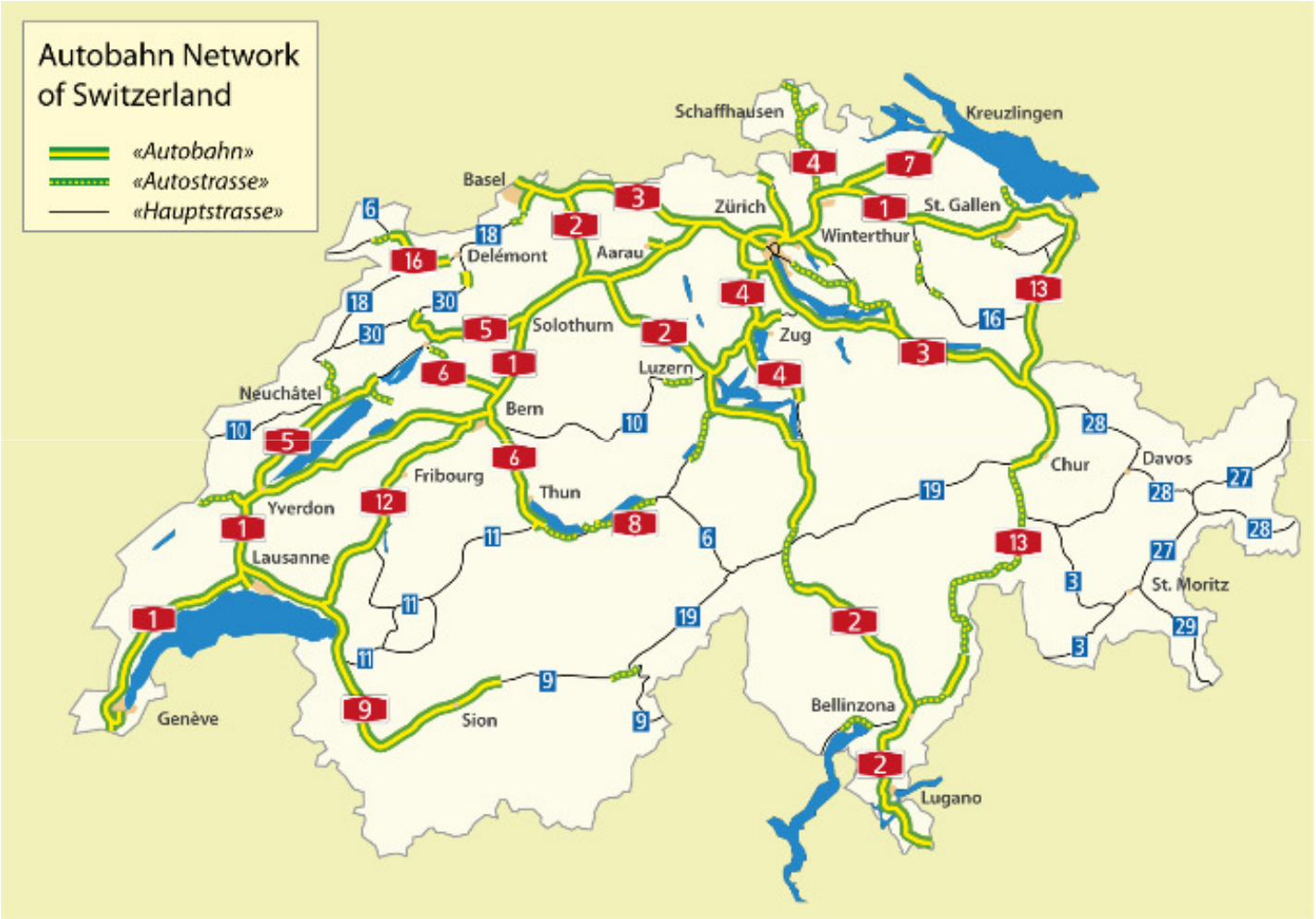
Exemple 1: Star Alliance Network



Star Alliance Network

http://www.staralliance.com/documents/20184/0/Global+Reach+Slide/

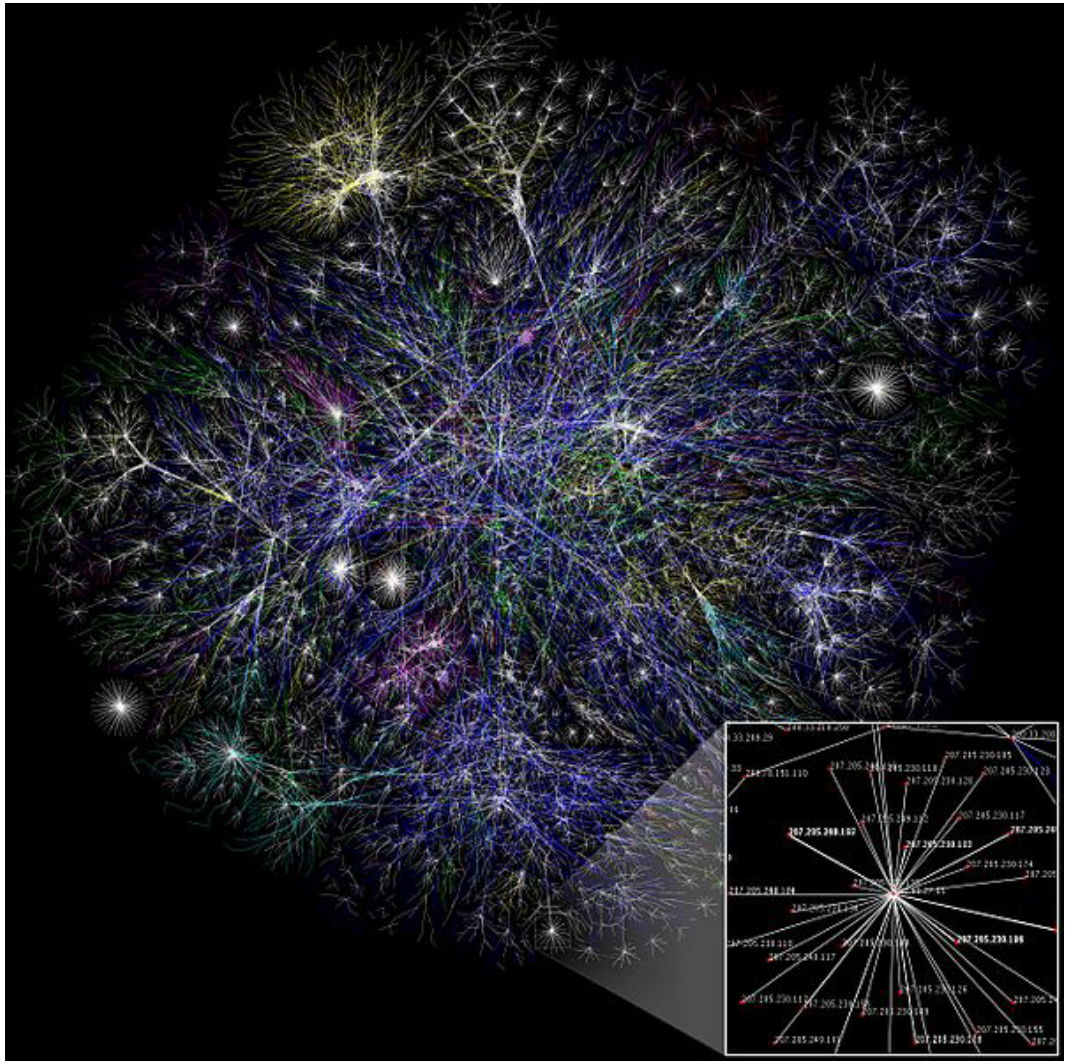
Exemple 2: Les Autoroutes Suisses



https://upload.wikimedia.org/wikipedia/commons/1/10/Image-Swiss-Highway-network-en.png

Exemple 3: Facebook Network

Exemple 4: Internet



https://en.wikipedia.org/wiki/Internet_Mapping_Project

Théorie des graphes

- La théorie des graphes est la branche des mathématiques qui étudies les propriétés de ces objets.
- La résolution de certains problèmes liés aux graphes fait appel à des algorithmes complexes.

Un article du mathématicien suisse Leonhard Euler publié en 1741, et traitant d'un problème appelé « les sept ponts de Königsberg » est considéré comme l'origine de la théorie des graphes.

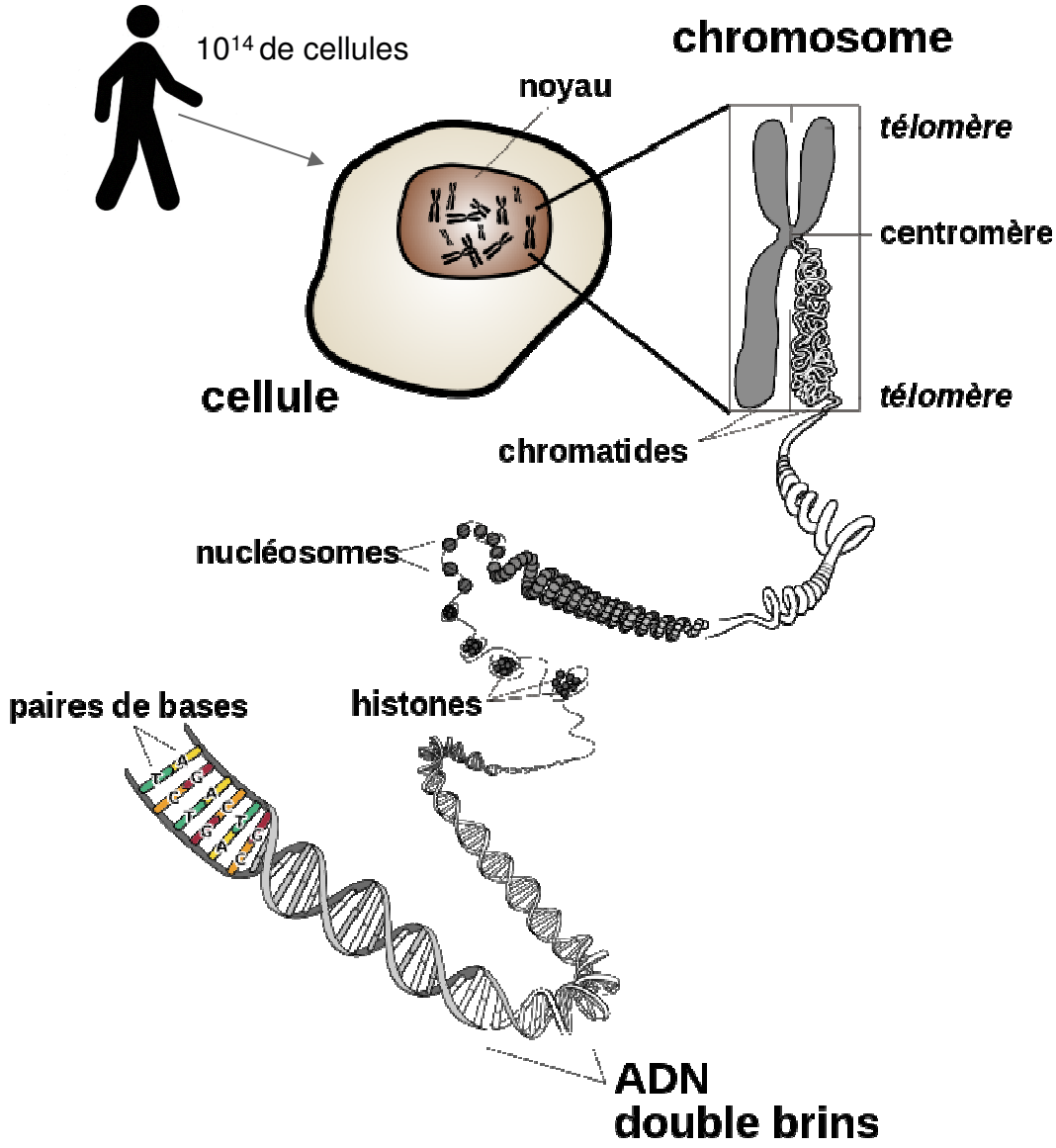
Le Dogme Central de la Biologie (1/2)

L'**ADN** est contenu dans le noyau de chacune des cellules de notre corps. Il est compose de 4 nucléotides **A**, **C**, **G**, **T**.

Le génome humain est composé de ~ **3'400'000'000** de ces nucléotides.

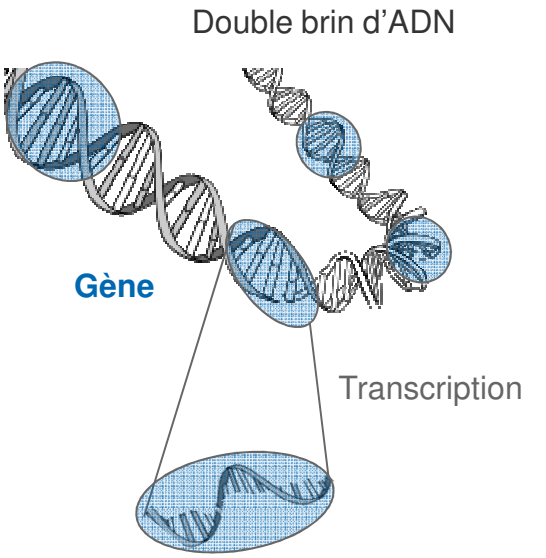
Arithmétique scolaire:

A raison de **4'000** caractères par page **A4**, il faudrait **850'000** pages soit un livre de **42.5** mètres de hauteur pour l'imprimer recto-verso.



https://commons.wikimedia.org/wiki/File:Chromosome_fr.svg

Le Dogme Central de la Biologie (2/2)



Traduction

Le double brin d'ADN contient des régions, appelées "gènes" qui contiennent les informations (encodées en utilisant l'alphabet {A,C,G,T}) pour créer des **protéines.**

Protéine

L'**ARNm** utilise les informations contenues dans l'ADN pour synthétiser des protéines.

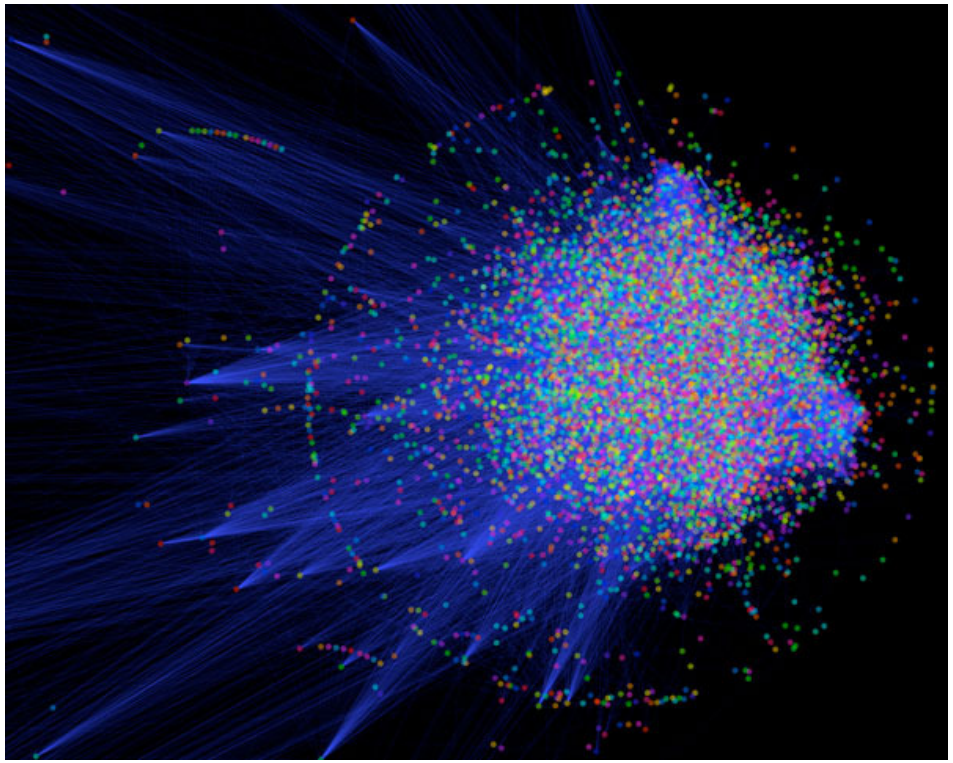
On estime à **26'517** le nombre de gènes codant pour des protéines.

Les protéines exécutent les tâches dans notre corps, Par exemple l'hémoglobine transporte l'oxygène. Les enzymes en particulier sont responsable de la biosynthèse des lipides.

Au-delà du Dogme: Réseaux biologiques

Les protéines interagissent entre elles en formant des complexes.

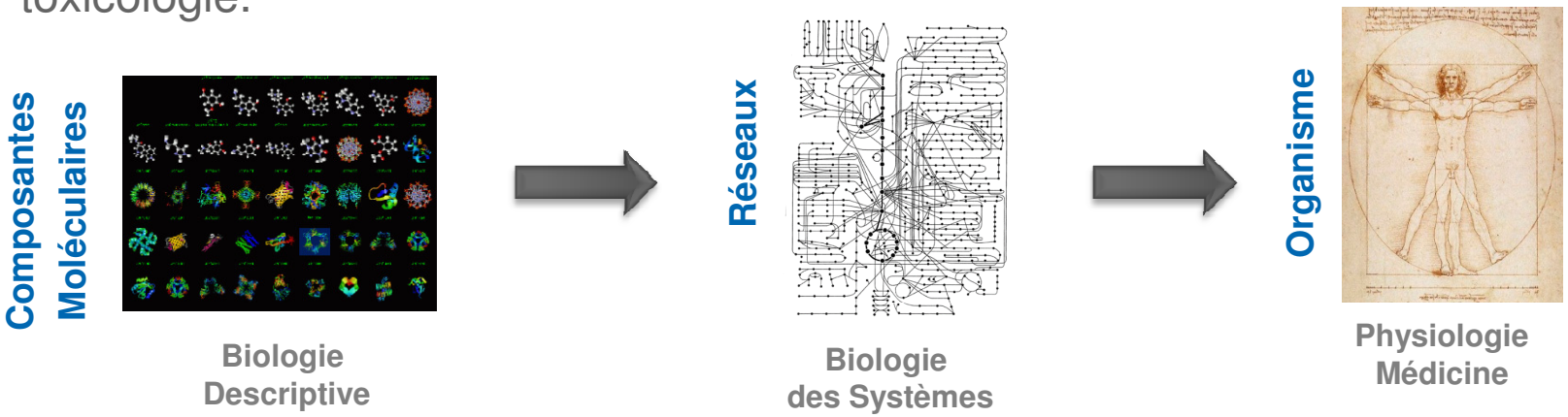
L'**interactome** humain est la description des interactions entre les protéines. C'est un **graphe** dont les sommets sont les protéines et dont les arêtes décrivent les interactions entre elles.



https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction

Réseaux Biologiques: Et c'est important?

La **biologie des systèmes** est une approche intégrative des différentes molécules (protéines, gènes,..) afin de décrire le système dans son ensemble.
La **toxicologie des systèmes** est l'application de la biologie des systèmes à la toxicologie.



Certaines maladies, comme le cancer, sont liées à la perturbation de réseaux biologiques.

CSH PRESS GENOME RESEARCH

Protein networks in disease

Trey Ideker[1] and Roded Sharan[2,3]

Opinion

Systems biology and the future of medicine

Joseph Loscalzo[1*] and Albert-Laszlo Barabasi[1,2]

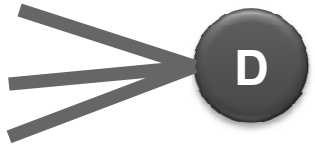
nature REVIEWS GENETICS

Network medicine: a network-based approach to human disease

Albert-László Barabási[*‡§], Natali Gulbahce*[*‡||] and Joseph Loscalzo[§]*

Hubs et Réseaux invariants d'échelle ("scale-free")

Dans un graphe, le **degré** d'un sommet est le nombre d'arêtes reliées à ce sommet.

Degré(

) = 3

On observe dans les réseaux biologiques et sociaux que certains sommets ont un degré bien plus élevé que la majorité des autres sommets: ce sont les **hubs.**

Hubs et Réseaux invariants d'échelle ("scale-free")
Réseau social

Lien entre les pages Wikipédia des chimistes, biologistes, physiciens et mathématiciens.

Les 3 "hubs" sont:

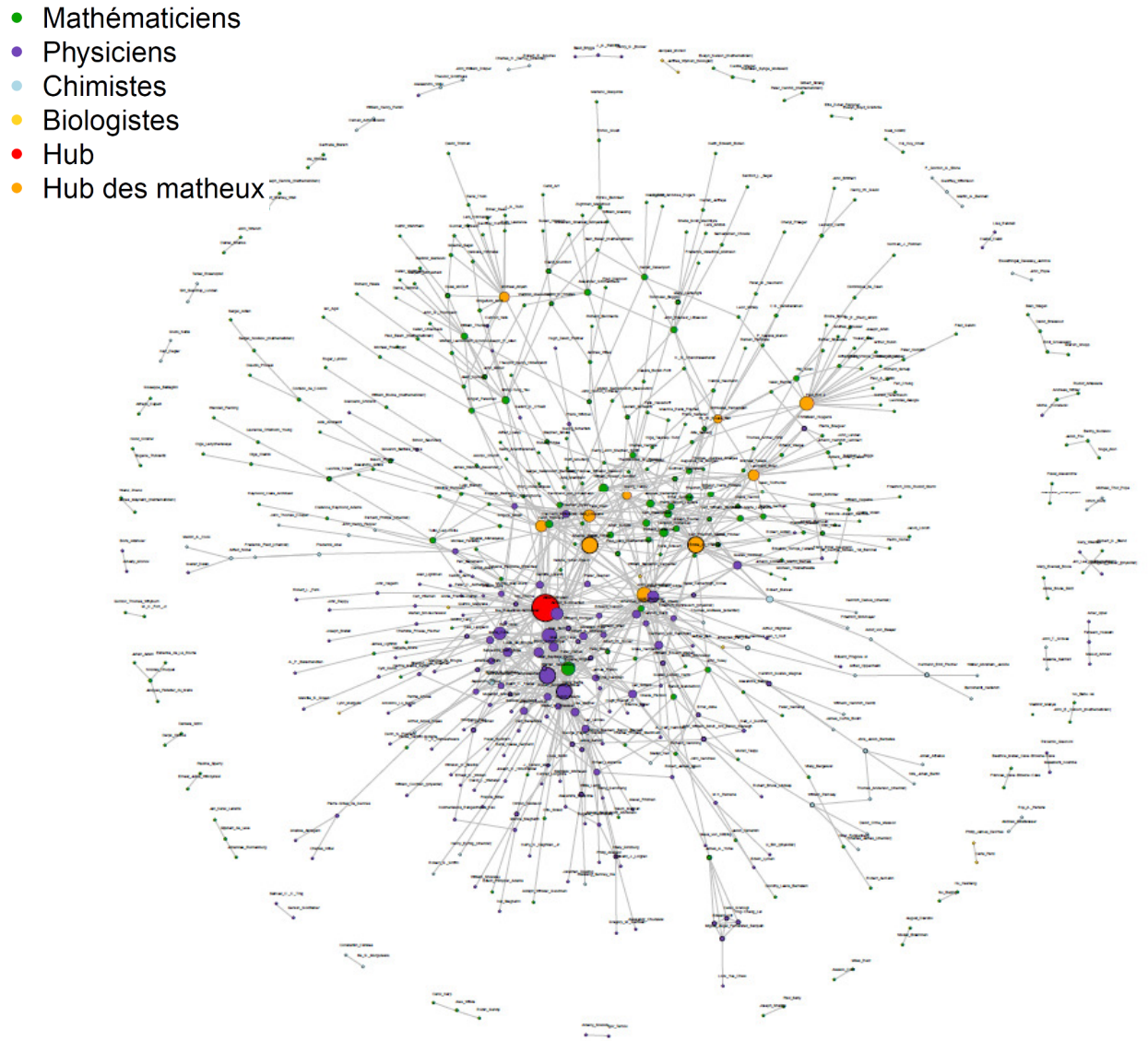
1. *Albert Einstein*
2. *Werner Heisenberg*
3. *David Hilbert*

Dans le sous-graphe des mathématiciens:

1. *Paul Erdös*
2. *Carl Friedrich Gauss*
3. *Felix Klein*

Et…

4. ***Leonhard Euler***



Parmi les 355 pages de biologistes connus, seulement 5 ont un lien vers d'autres scientifiques!

Hubs et Réseaux invariants d'échelle ("scale-free")
Réseau social

Lien entre les pages Wikipédia des chimistes, biologistes, physiciens et mathématiciens.

Les 3 "hubs" sont:

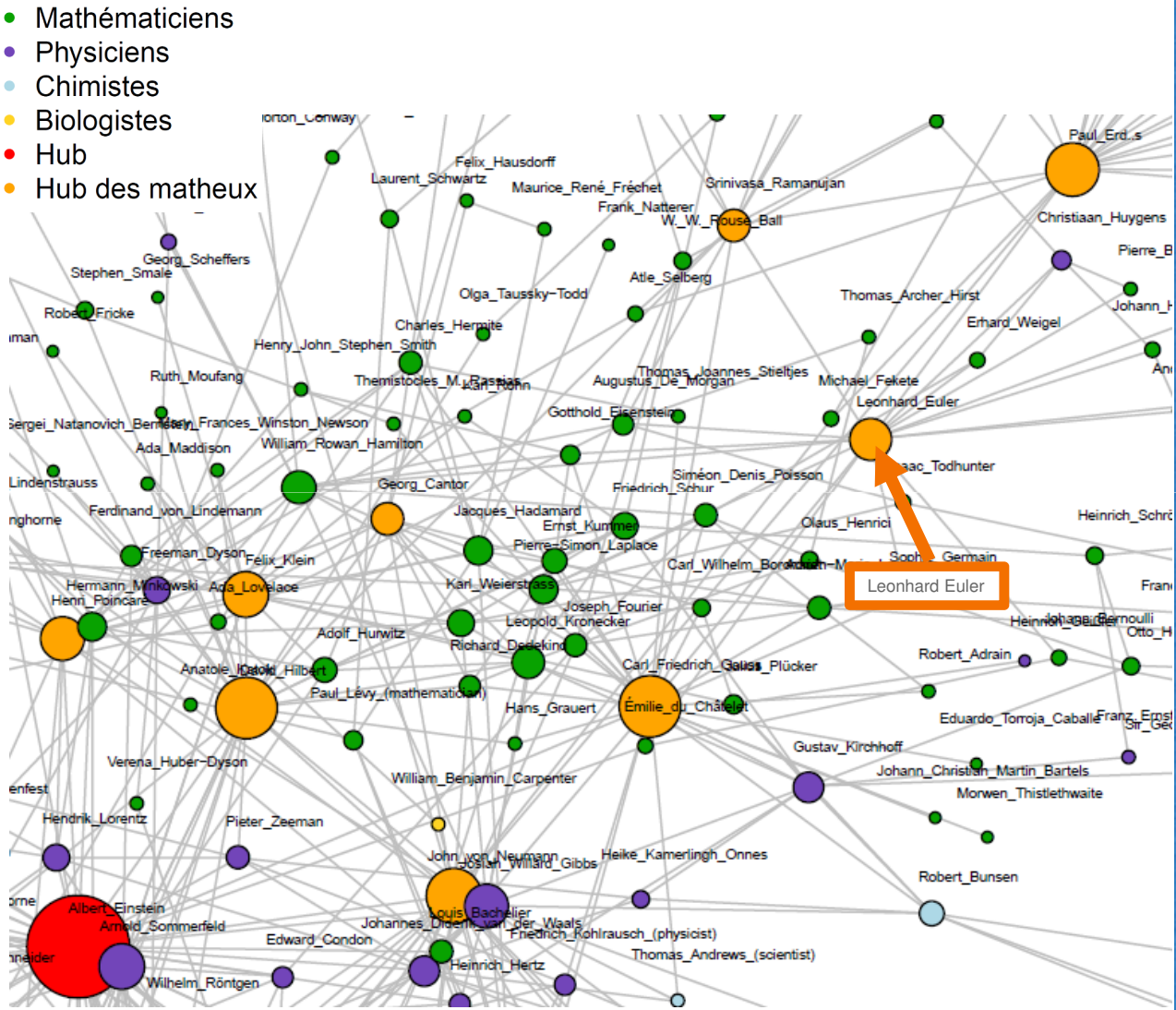
1. *Albert Einstein*
2. *Werner Heisenberg*
3. *David Hilbert*

Dans le sous-graphe des mathématiciens:

1. *Paul Erdös*
2. *Carl Friedrich Gauss*
3. *Felix Klein*

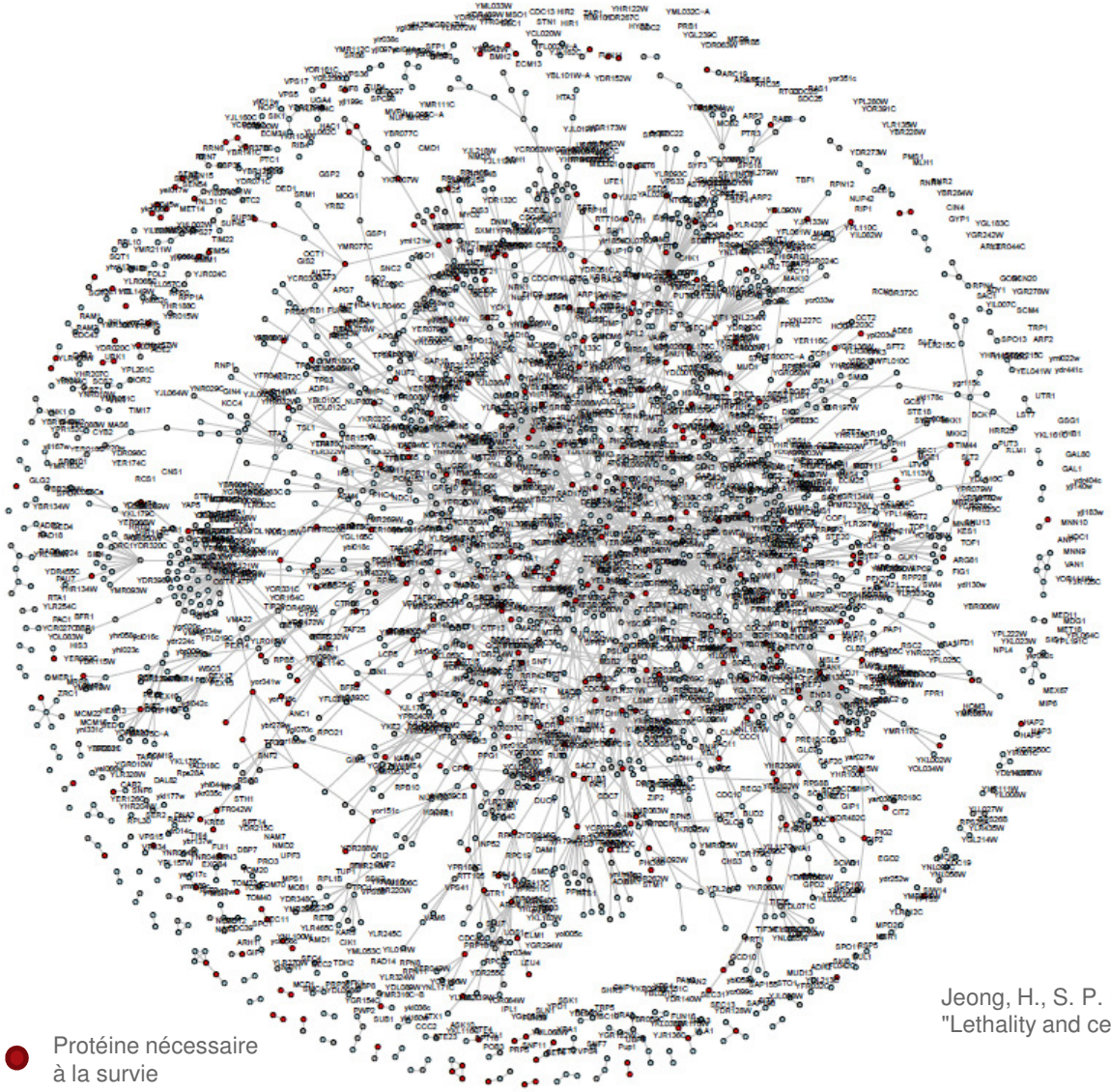
Et…

4. ***Leonhard Euler***



Parmi les 355 pages de biologistes connus, seulement 8 ont un lien vers d'autres scientifiques!

Hubs et Réseaux invariants d'échelle ("scale-free")
Réseau biologique



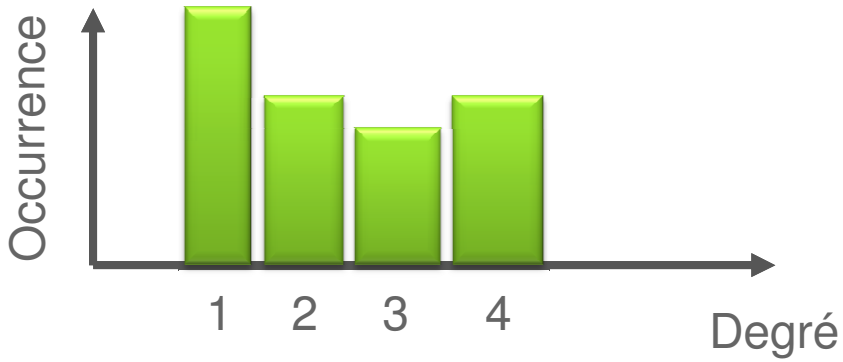
Protéine nécessaire à la survie

Interactions protéine-protéine dans la levure. Le degré des protéines nécessaires à la survie est plus grand (en moyenne de 49%)!

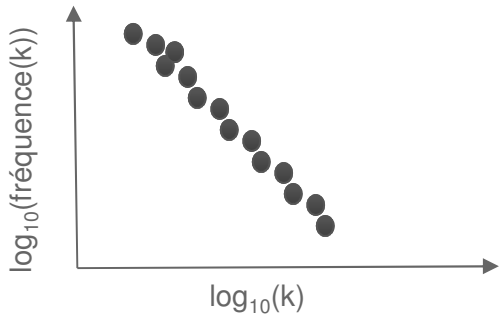
Jeong, H., S. P. Mason, A.-L. Barabasi and Z. N. Oltvai. (2001). "Lethality and centrality in protein networks." Nature 411(6833): 41-42.

Hubs et Réseaux invariants d'échelle ("scale-free")

On peut alors s'intéresser à la distribution des degrés dans un graphe:

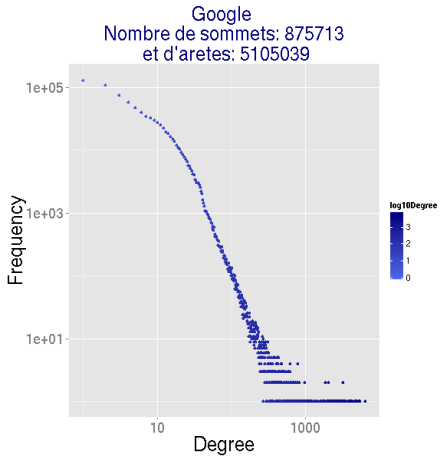


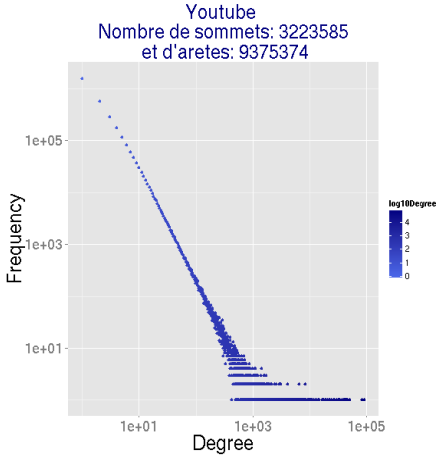
Un réseau est invariant d'échelle («scale-free») si la distribution des degrés de ses sommet suit une loi exponentielle: $P(k) \sim k^{-c}$
Donc, pour détecter une telle propriété, il suffit de représenter le $\log_{10}(k)$ vs. $\log_{10}(\text{fréquence}(k))$: si les points sont alignés, le réseau est invariant d'échelle!

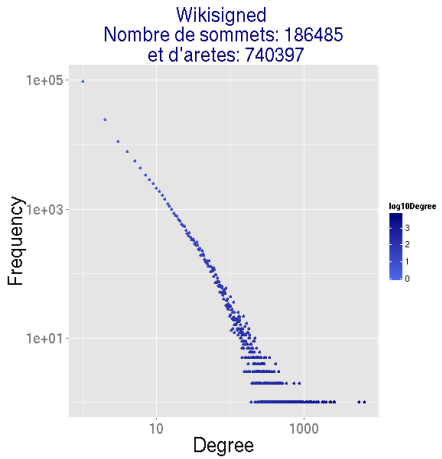


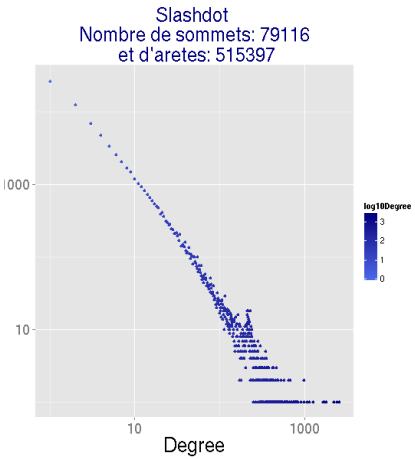
Invariance d'échelle: Une propriété communément partagée par les réseaux biologiques et sociaux

Sociologie

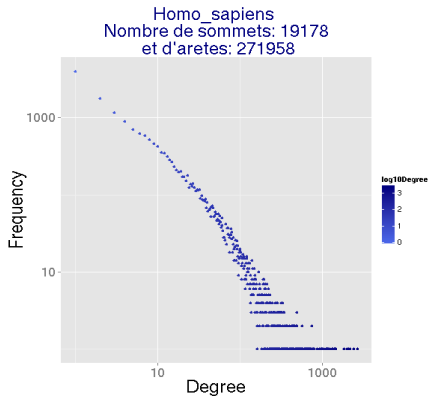


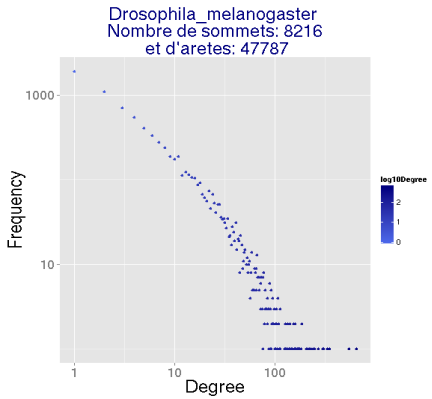


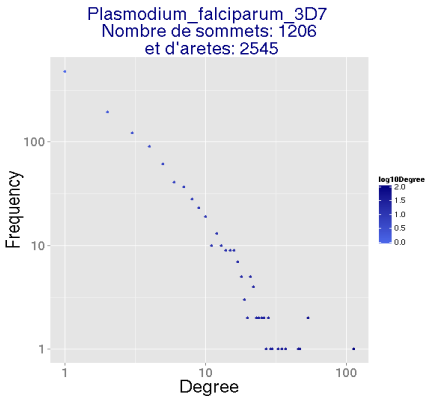


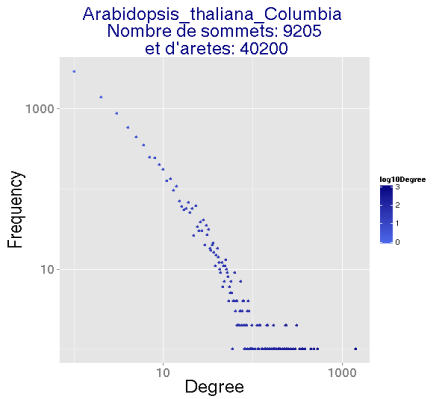


Biologie





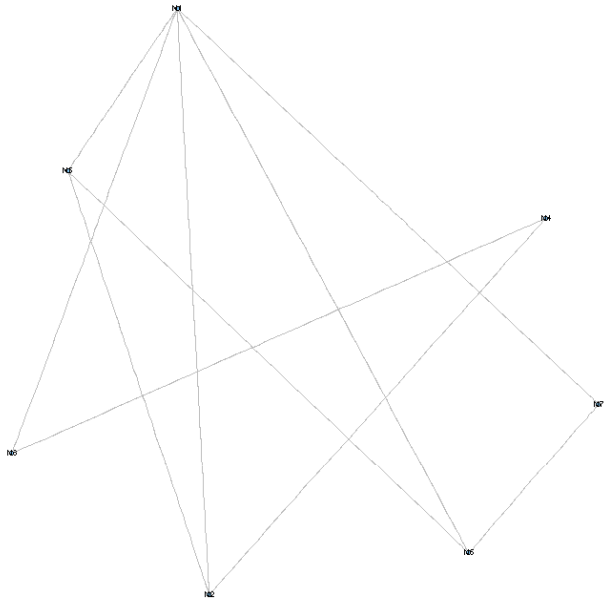


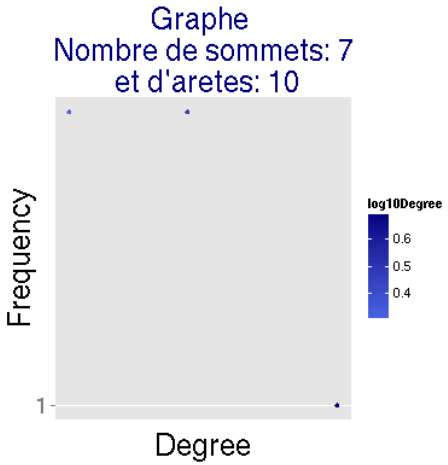


Hubs et Réseaux invariants d'échelle ("scale-free")

Peut-on essayer d'expliquer cette propriété?

On utilise le modèle aléatoire de l'**attachement préférentiel** pour décrire l'évolution d'un réseau: un nouveau membre du réseau social (ou une nouvelle molécule en biologie?) aura tendance à se mettre en relation avec quelqu'un qui a déjà beaucoup de relations («key opinion leaders», «personnalités connues», etc...)





Communautés dans les réseaux

- **Définition**: Il existe dans les réseaux des régions où les sommets sont densément reliés entre eux. Ces sous-réseaux sont appelés **communautés.**

- Détecter les communautés dans un graphe n'est pas chose aisée. Certains algorithmes pour représenter les graphes essaient de les mettre en évidence; mais cette approche n'est généralement pas applicable.

- Des méthodes spécifiques sont nécessaires. Une méthode mathématique pour détecter les communautés: le **clustering spectral.**

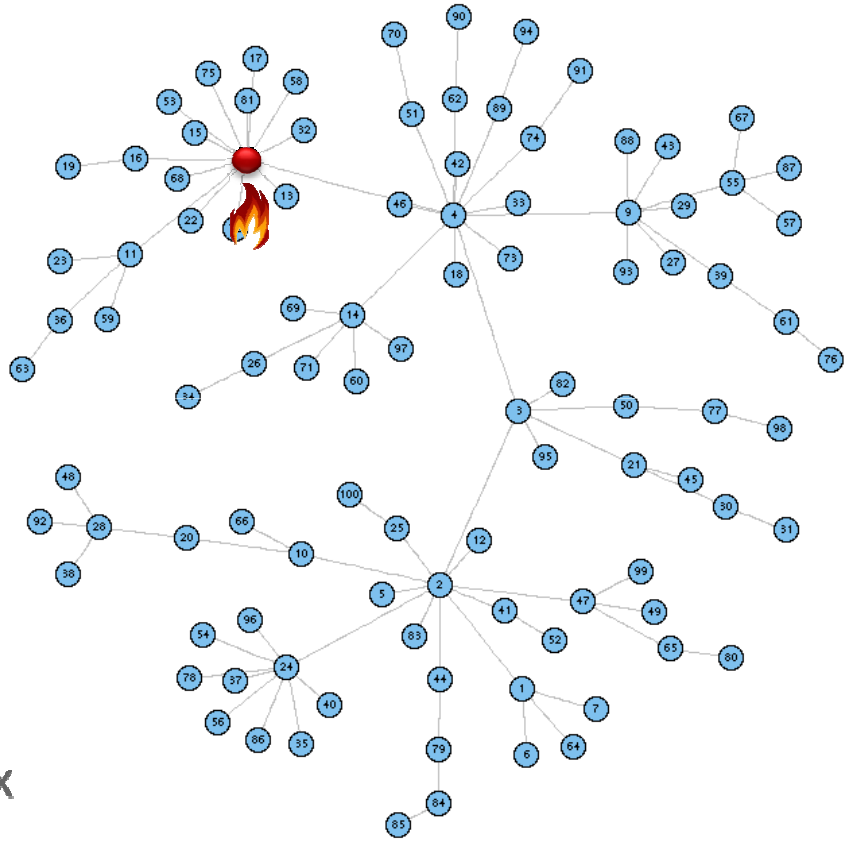
Communautés dans les réseaux: Le Laplacien d'un graphe

- Petite expérience de l'esprit: Considérons un graphe en fil de fer. Chauffons un sommet avec un briquet: la chaleur va se diffuser dans le graphe: c'est l'équation de la chaleur:

$$\frac{\partial f}{\partial t} = -\Delta f$$

où f décrit l'évolution de la chaleur.

- Condition initiales en x_0: $f(x_0,0)=900°C$ et $f(x,0)=22°C$ si x est différent de x_0.



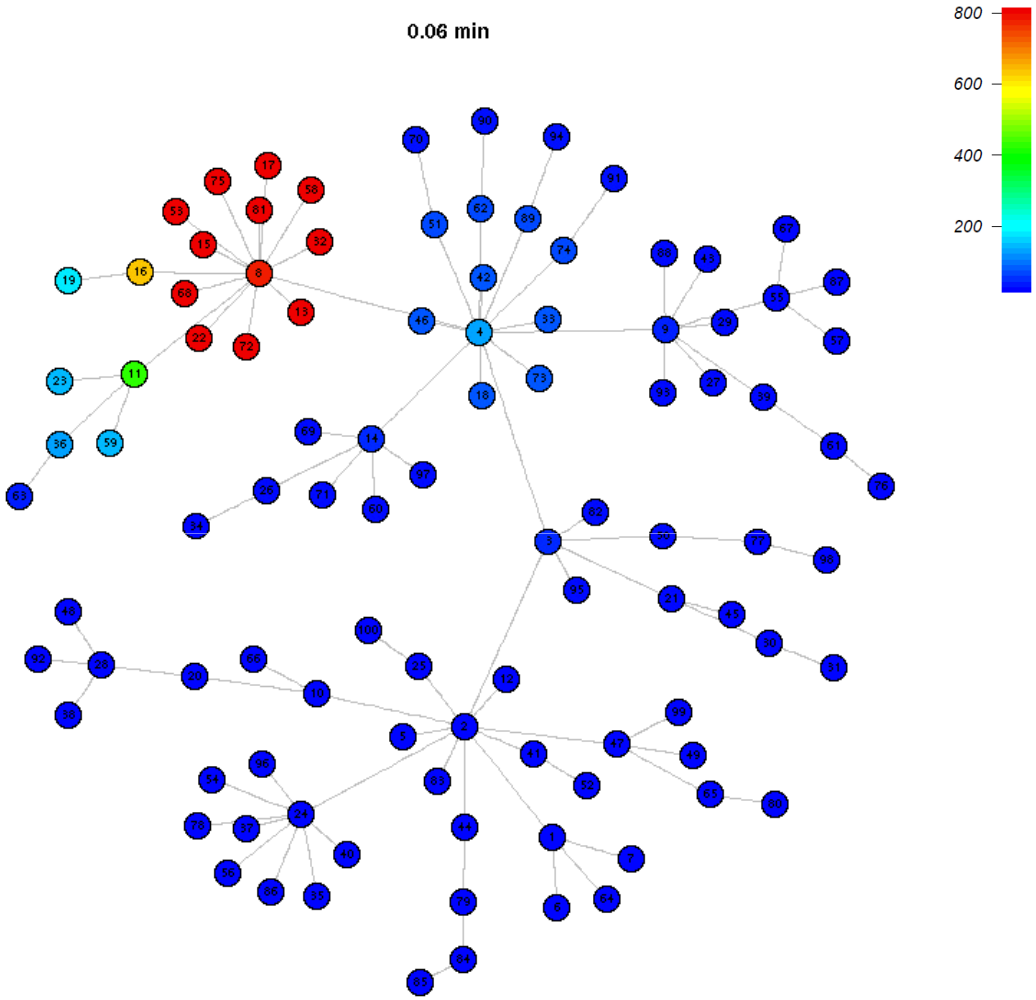
Δ est appelé le **Laplacien** du graphe, habituellement dénotée L.

Communautés dans les réseaux: Le Laplacien d'un graphe

L'équation de la chaleur et ses solutions permettent d'identifier des régions du graphe ou la chaleur se propage facilement; et donc d'identifier des communautés.

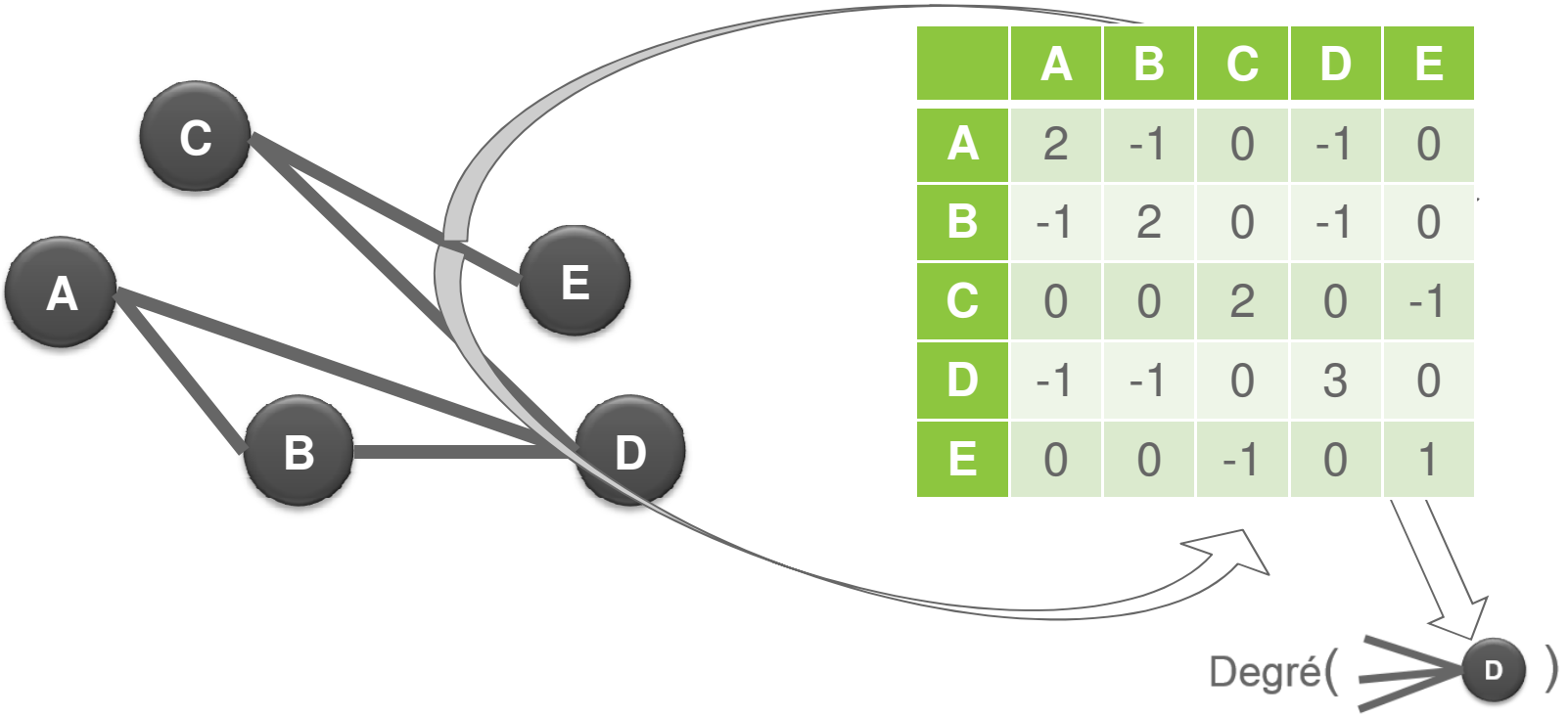
Cette information réside principalement dans le Laplacien.

C'est pourquoi, l'étude (spectrale) de cette matrice est utilisée.



Le Laplacien d'un graphe

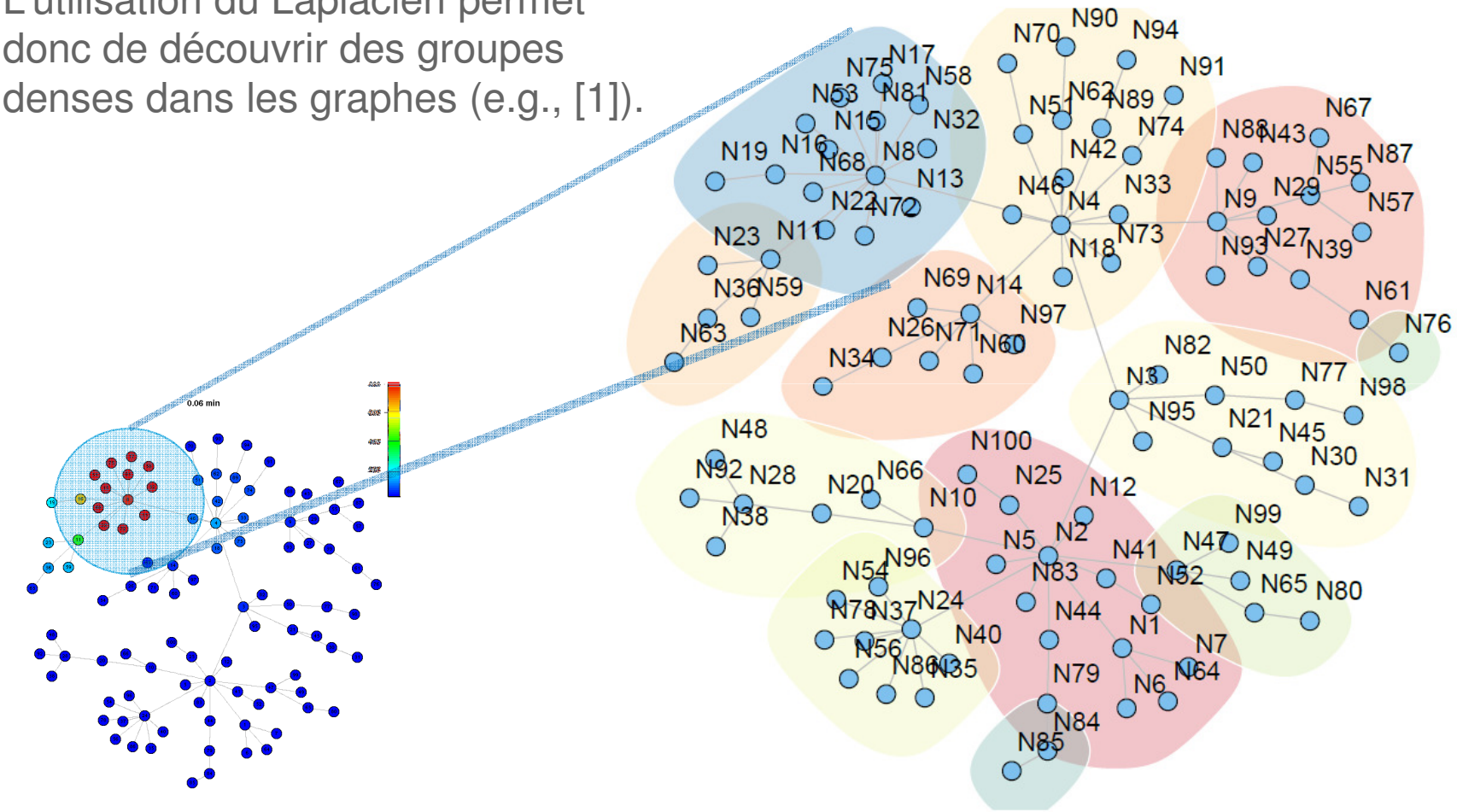
- L est très facile à construire:



	A	B	C	D	E
A	2	-1	0	-1	0
B	-1	2	0	-1	0
C	0	0	2	0	-1
D	-1	-1	0	3	0
E	0	0	-1	0	1

Communautés dans les réseaux: Exemples

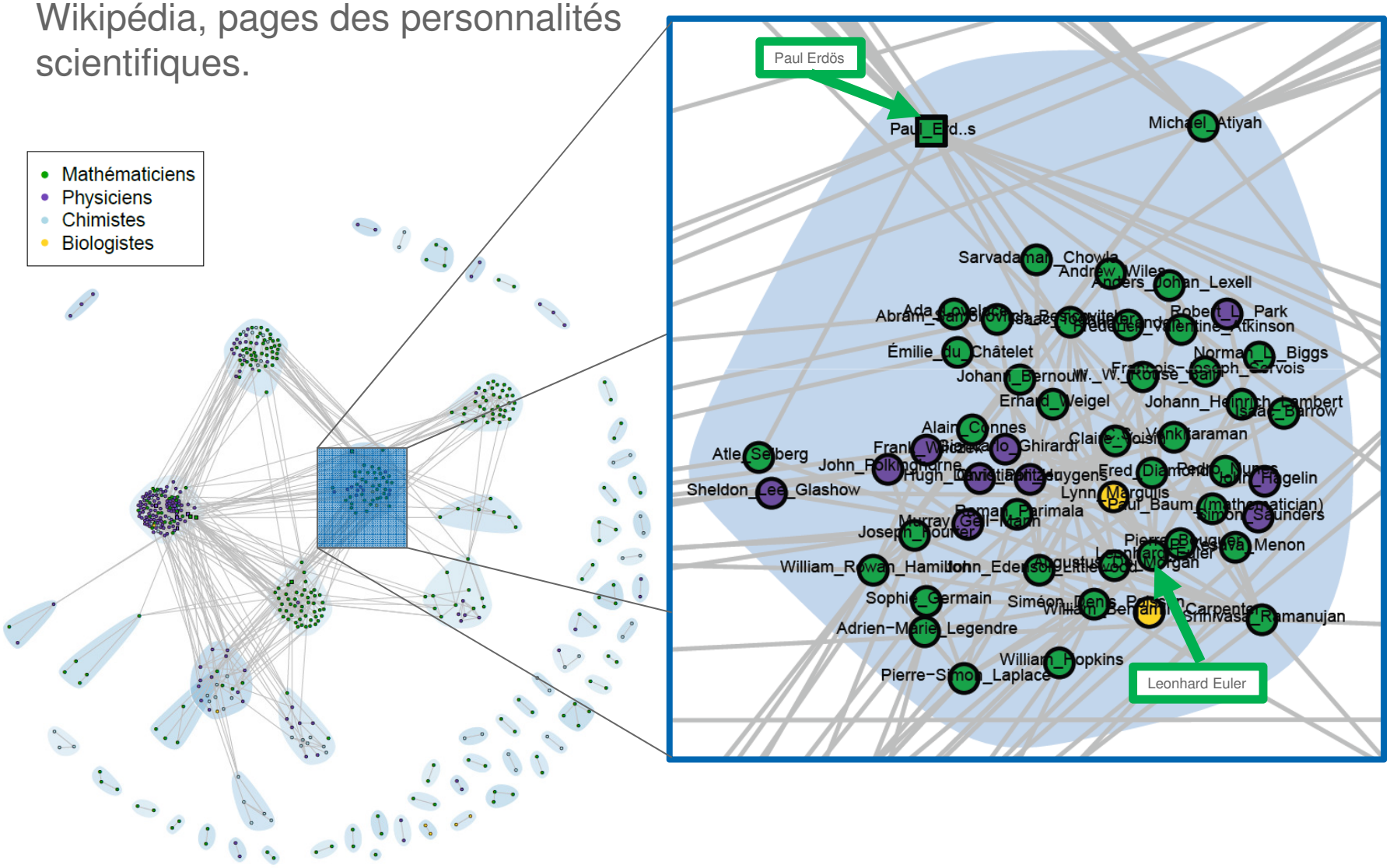
L'utilisation du Laplacien permet donc de découvrir des groupes denses dans les graphes (e.g., [1]).



[1] MEJ Newman: "Finding community structure using the eigenvectors of matrices", Physical Review E 74 036104, 2006.

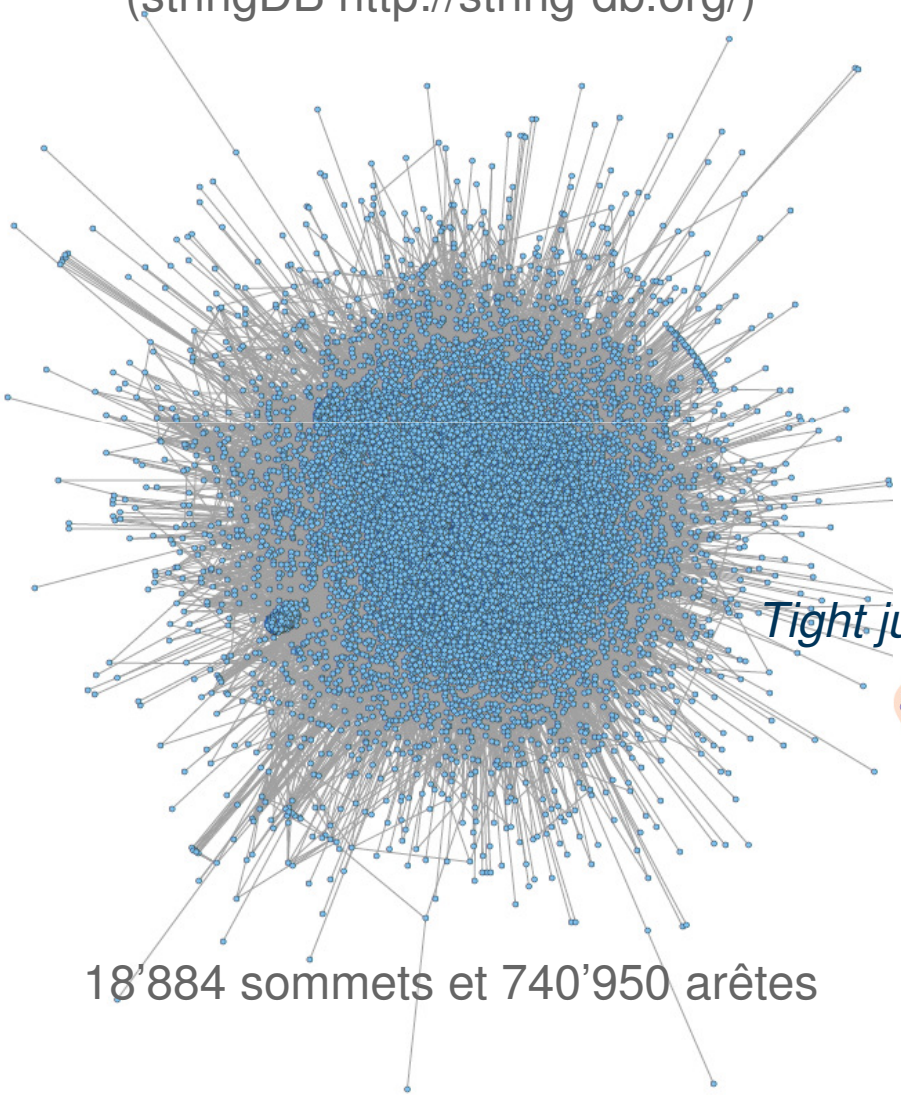
Communautés dans les réseaux: Réseau social

Wikipédia, pages des personnalités scientifiques.



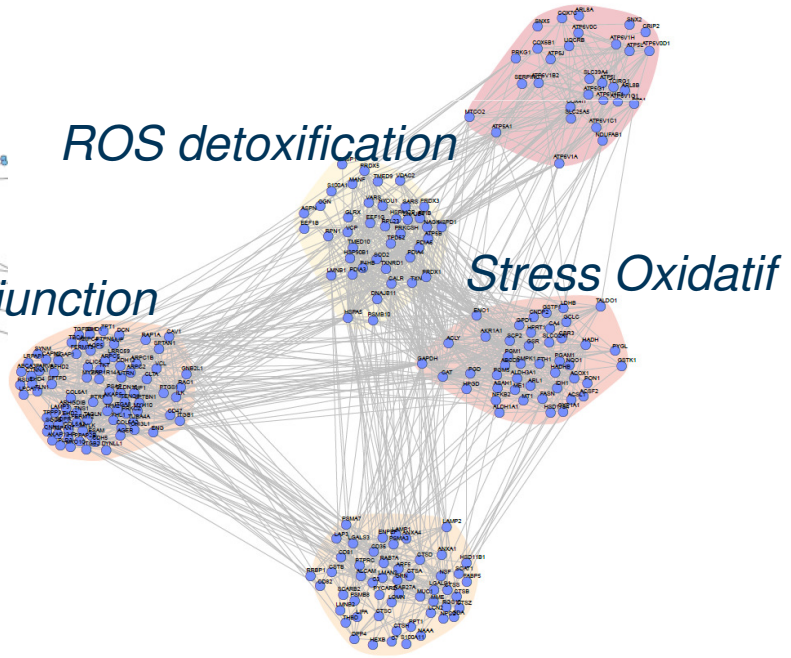
Communautés dans les réseaux: Réseau biologique

Réseau biologique
(stringDB http://string-db.org/)



18’884 sommets et 740’950 arêtes

L’étude des protéines répondant à un stimulus donné se groupent en communautés décrivant des processus biologiques.



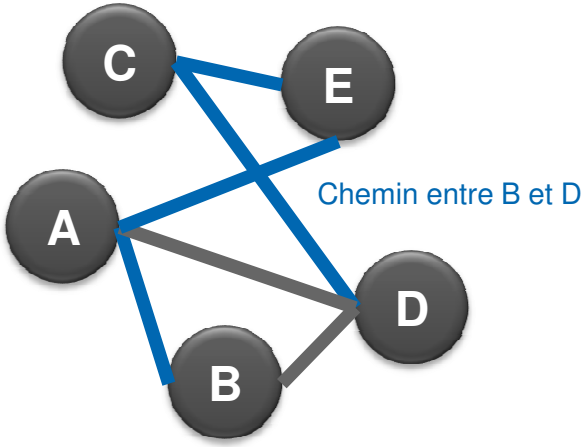
Frustration d'un graphe signé

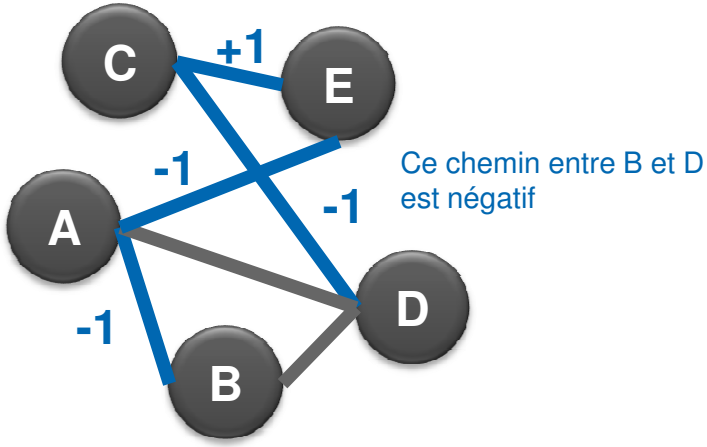
- "*Les amis de mes amis sont mes amis et les ennemis de mes amis sont mes ennemis*". Cette citation bien connue reflète une propriété des réseaux sociaux.

- Problème du message: Si les sommets sont des personnes et que les arêtes sont signées (relations amicales (+1) ou inamicales (-1)), on cherche alors a communiquer un message.
  - Paul communique le message à un de ses voisins. Si la relation est amicale, le message est transmis tel quel; sinon c'est le contraire de ce message qui est communiqué.
  - Si Paul reçoit en retour l'information; recevra-t-il le message d'origine ou son contraire?

Frustration d'un graphe signé

- Un chemin dans un graphe est une suite d'arêtes partageant 2 à 2 un sommet. Si le graphe est signé, le signe d'un chemin est le produit des signes de ces arêtes. Un cycle est un chemin finissant à son point de départ.

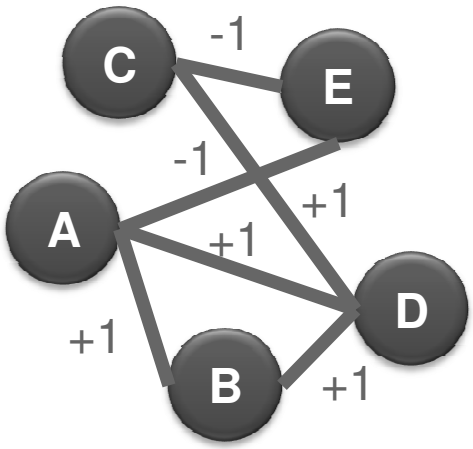




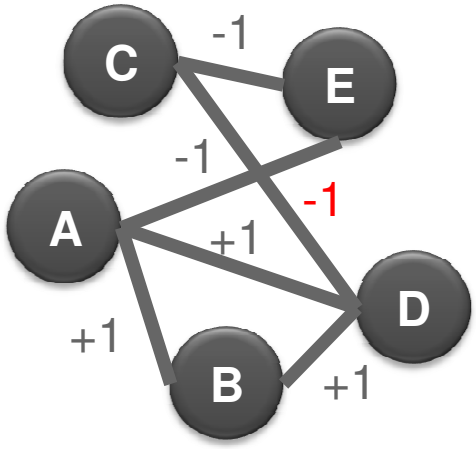
- Le problème est de savoir si le graphe signé sous-jacent est **équilibré**: *un graphe signé est équilibré si tous ces cycles sont positifs*.

Frustration d'un graphe signé

- Donc si le graphe est équilibré, Paul recevra le message d'origine!



Graphe équilibré



Graphe pas équilibré

Définition: Le nombre d'arêtes dont il faut changer le signe (où enlever) pour rendre un graphe signé équilibré est appelé **frustration.**

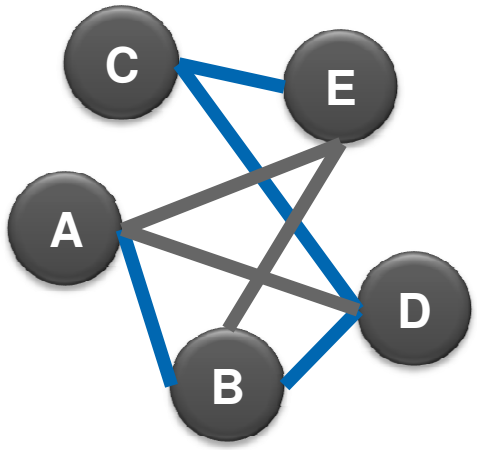
Donc un graphe est équilibré si et seulement si sa frustration est nulle.

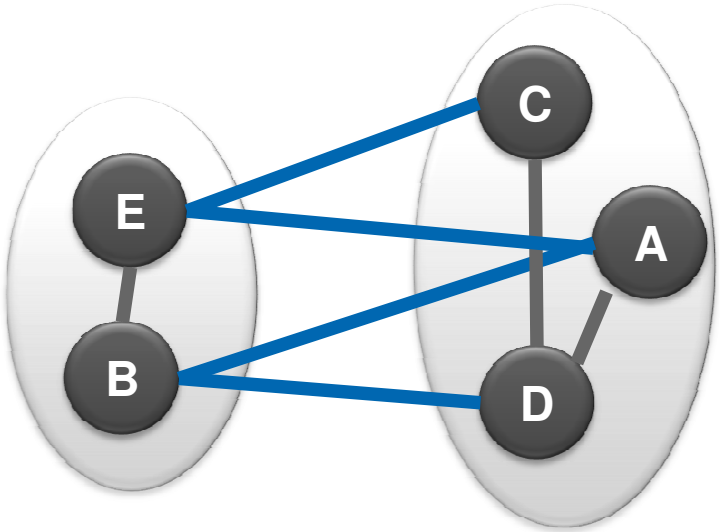
Frustration d'un graphe signé

- On peut montrer assez facilement:

Théorème:

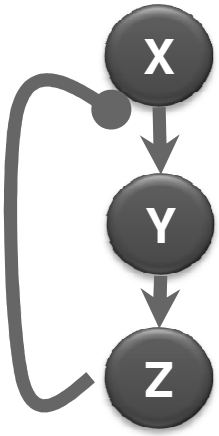
Un graphe signé est équilibré si et seulement si on peut partitionner ses sommets en deux groupes opposés d'amis (arêtes positives entre les sommets d'un groupe) et uniquement des arêtes négatives reliant les deux groupes: c'est-à-dire deux alliances ennemies!





-1
+1

Frustration et Réseaux biologiques: Feedback négatif



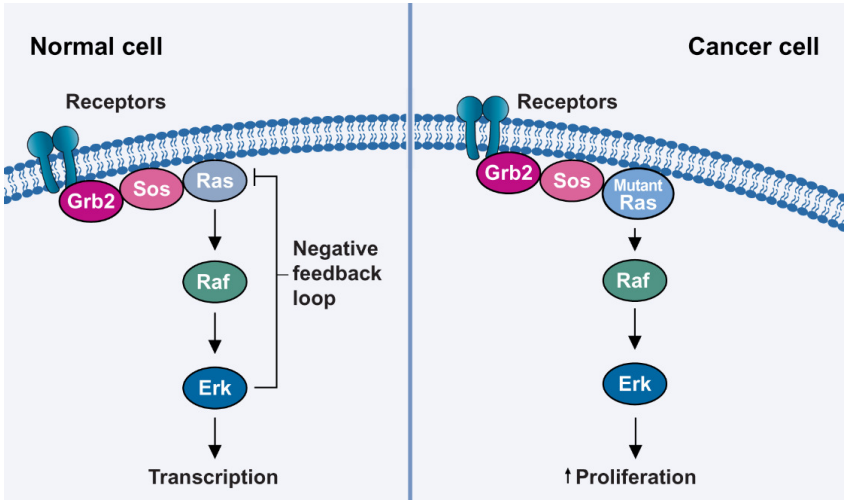
Un **modèle mathématique** décrivant ce système dynamique non linéaire:

$$\frac{dx}{dt} = k_1 - k_1 x(t) + \frac{k_3}{(k_4^n + z(t)^n)}$$

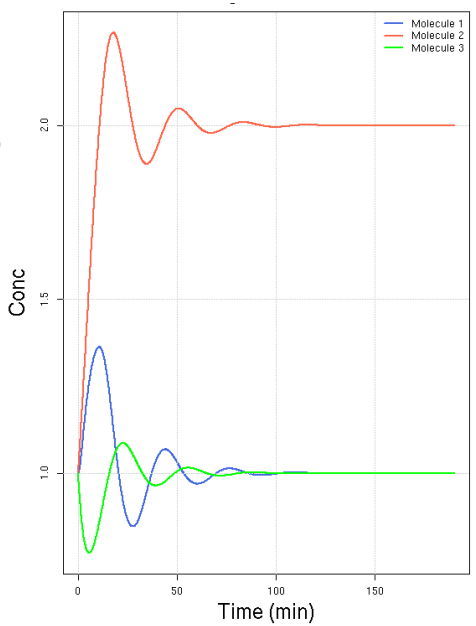
$$\frac{dy}{dt} = k_5\, x(t) - k_6 y(t)$$

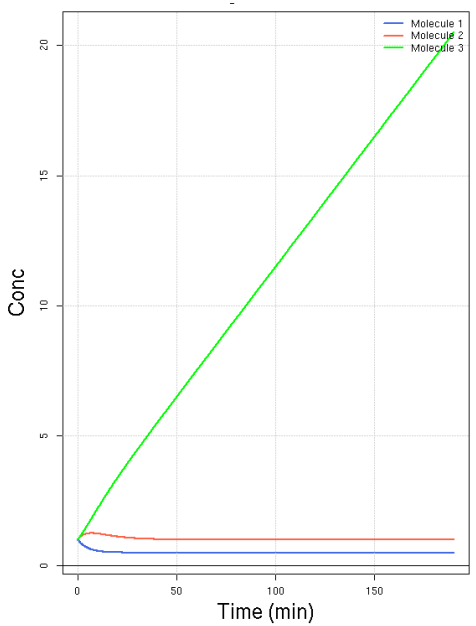
$$\frac{dz}{dt} = k_7\, y(t) - k_8 z(t)$$

k_1=0.1, k_2=0.2, k_3=0.2, k_4=1,
k_5=0.2, k_6=0.1, k_7=0.1, k_8=0.2,
n=12,x(0)=1, y(0)=1, z(0)=1

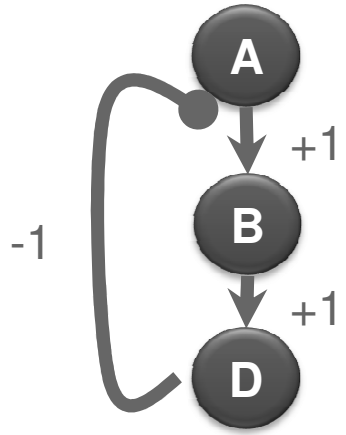


http://www.biooncology.com/molecular-causes-of-cancer/proliferative-signaling

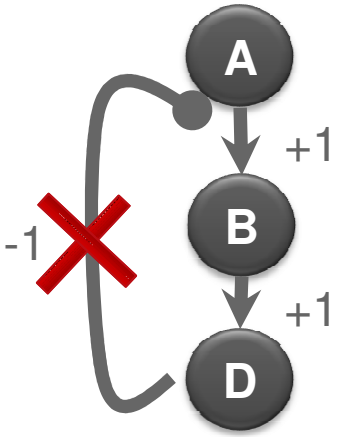




Frustration et Réseaux biologiques: Feedback négatif



Frustration=1, ce réseau n'est pas équilibré.

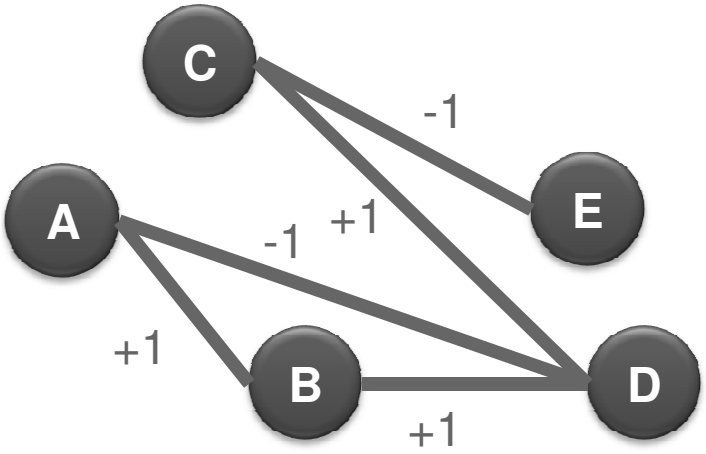


Frustration=0, ce réseau est équilibré.

Dans les réseaux biologiques, la frustration indique le degré d'auto-régulation d'un système. Si il n'y a pas de frustration (i.e., le réseau est équilibré), le système dynamique sous-jacent devient monotone.

Le Laplacien d'un graphe signé

- C'est une simple adaptation du Laplacien déjà décrit:



	A	B	C	D	E
A	2	-1	0	**1**	0
B	-1	2	0	-1	0
C	0	0	2	0	**1**
D	**1**	-1	0	3	0
E	0	0	**1**	0	1

Frustration: Comment calculer ce nombre et identifier un ensemble d'arêtes?

Deux théorèmes faisant le lien entre la frustration et le Laplacien signé:

Théorème 1: Soit $\mathcal{F}(G_\sigma)$ la frustration d'un graphe signé connexe G_σ et soit $\lambda_1(G_\sigma)$ la première valeur propre de L_σ. Alors:

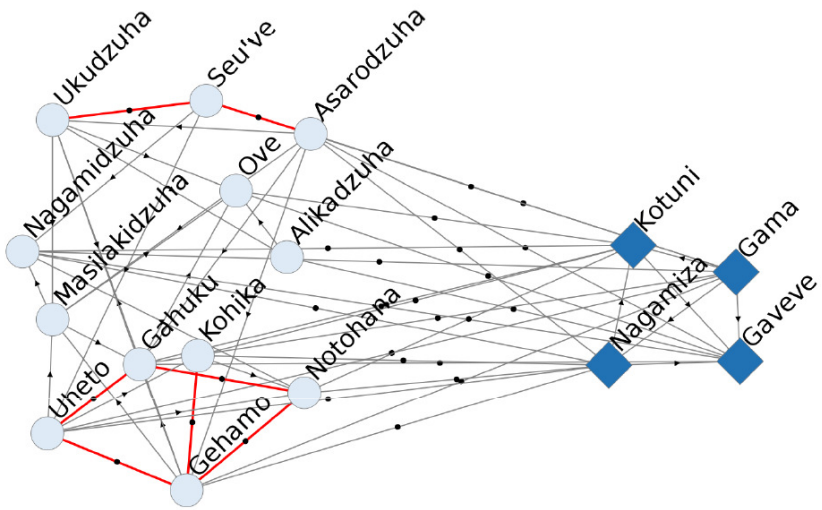
$$\frac{n}{4} \cdot \lambda_1(G_\sigma) \leq \mathcal{F}(G_\sigma) \leq \frac{n}{\sqrt{2}} \cdot \sqrt{\lambda_1(G_\sigma)(2\Delta - \lambda_1(G_\sigma))}$$

où Δ est le degré maximal du graphe.

Théorème 2: On a $$\mathcal{F}(G_\sigma) = \frac{n}{2} \min_{\substack{f \neq 0 \in l^2(\hat{V}), \ f \\ symmetric}} \frac{\sum_{x \sim y} |f(x) - f(y)|}{\sum_x |f(x)|}$$

Trouver une solution au théorème 2 permet alors d'identifier effectivement un ensemble d'arêtes dont la suppression rend le graphe équilibré.

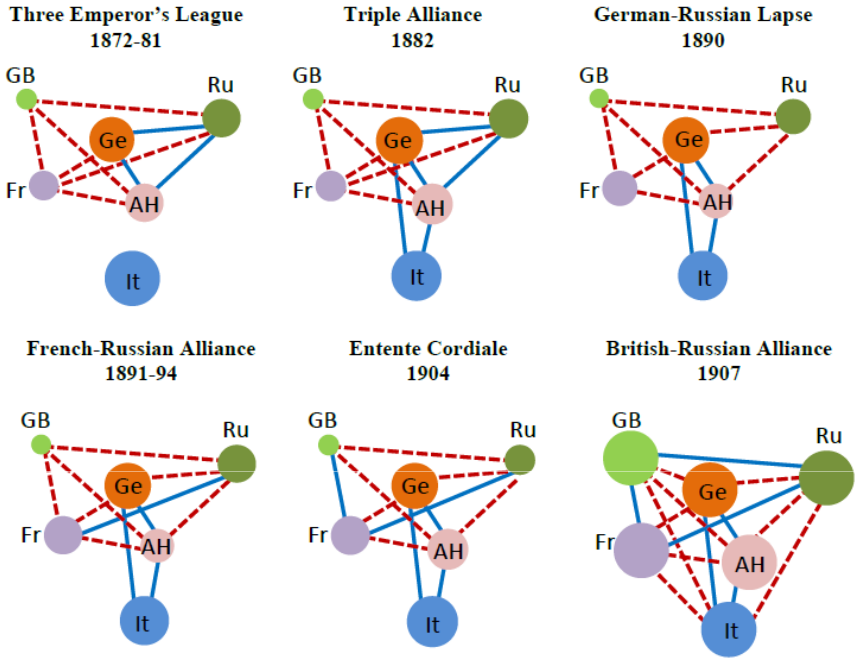
Frustration et Réseaux sociaux



Le premier exemple est un réseau social bien connu qui décrit les relations entre les tribus des hauts plateaux, en Nouvelle-Guinée dans les années 50.

Frustration=7

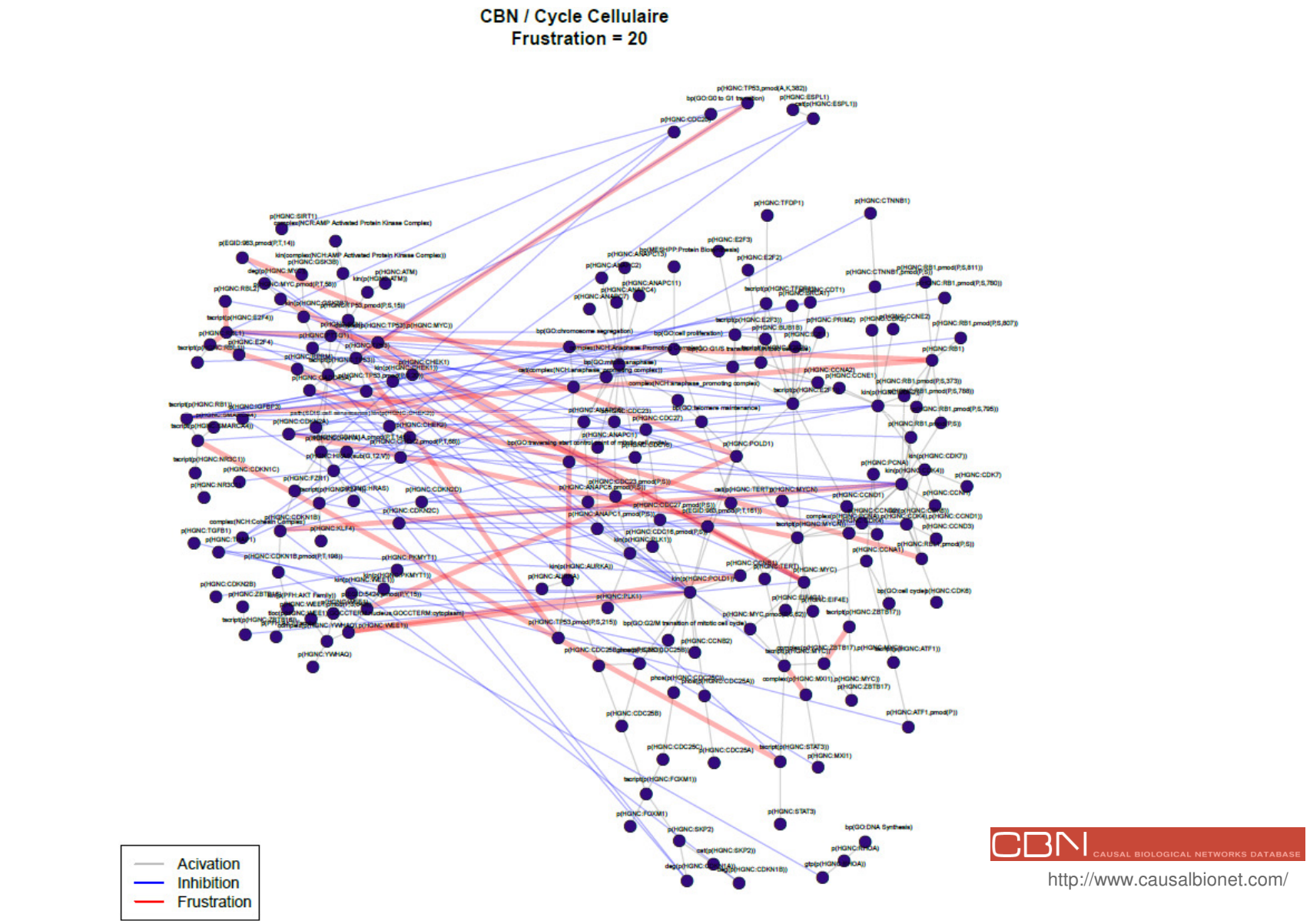
Kenneth E Read. “Cultures of the central highlands, new guinea.” *Southwestern Journal of Anthropology*, pages 1–43, 1954.



Les réseaux d’alliance, ici les relations entre l’Allemagne (Ge), l’empire Austro-Hongrois (AH), l’Italie (It), l’Angleterre (GB), la France (Fr) et la Russie (Ru). Les graphes ont tendance à devenir équilibrés: Frustration en 1907 = 0.

http://arxiv.org/ftp/arxiv/papers/1406/1406.2132.pdf

Frustration: Exemple plus complexe le cycle cellulaire



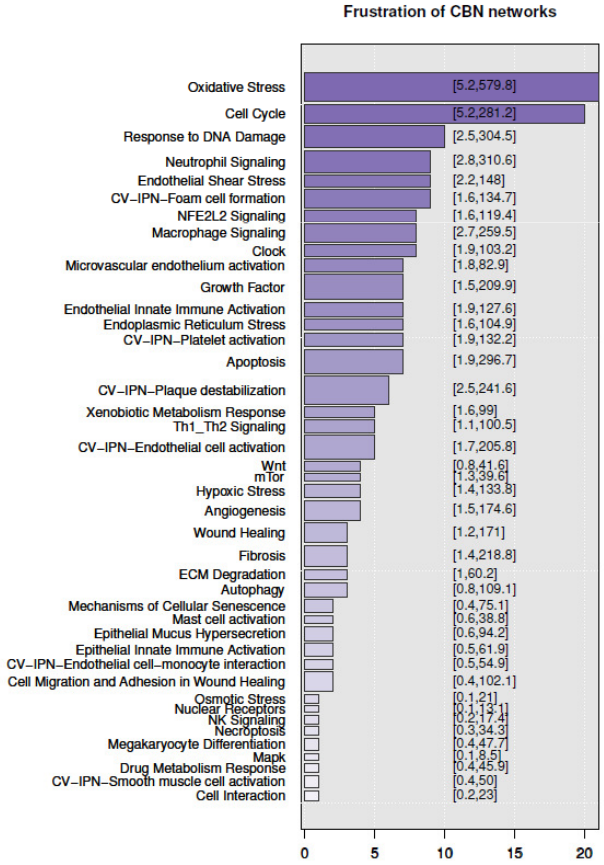
CBN CAUSAL BIOLOGICAL NETWORKS DATABASE

http://www.causalbionet.com/

PMI RESEARCH & DEVELOPMENT

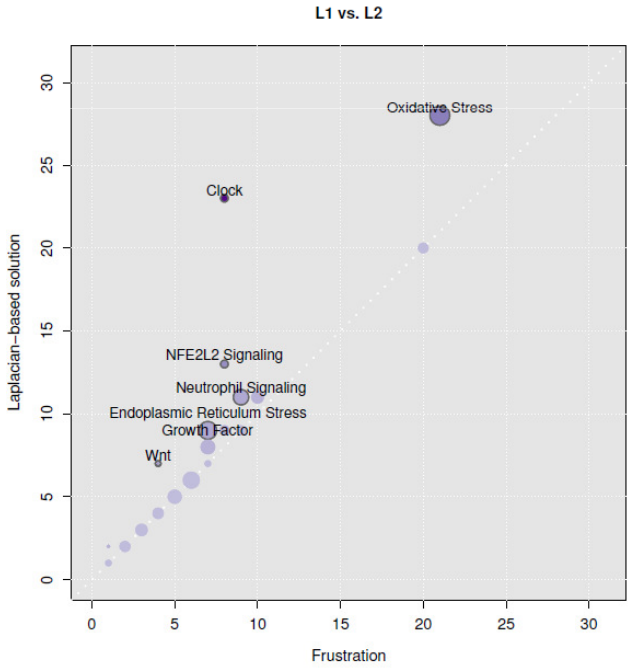
Le Laplacien est un bon indicateur de la frustration.

Si on estime la frustration basé sur la première fonction propre de L_σ (Théorème 1) ou si on la calcule en utilisant le théorème 2, les deux approches s'accordent à quelques exceptions près. Les réseaux de la base de donnée CBN servent d'exemple ici.



CBN CAUSAL BIOLOGICAL NETWORKS DATABASE

http://www.causalbionet.com/



Et les 7 ponts alors!?

- *Le problème consiste à déterminer s'il existe ou non une promenade dans les rues de Königsberg permettant, à partir d'un point de départ au choix, de passer une et une seule fois par chaque pont, et de revenir à son point de départ, étant entendu qu'on ne peut traverser le fleuve qu'en passant sur les ponts.*

Théorème d'Euler (1736):

Un graphe a un cycle eulérien si et seulement si chaque sommet a un degré pair.

Merci de votre attention