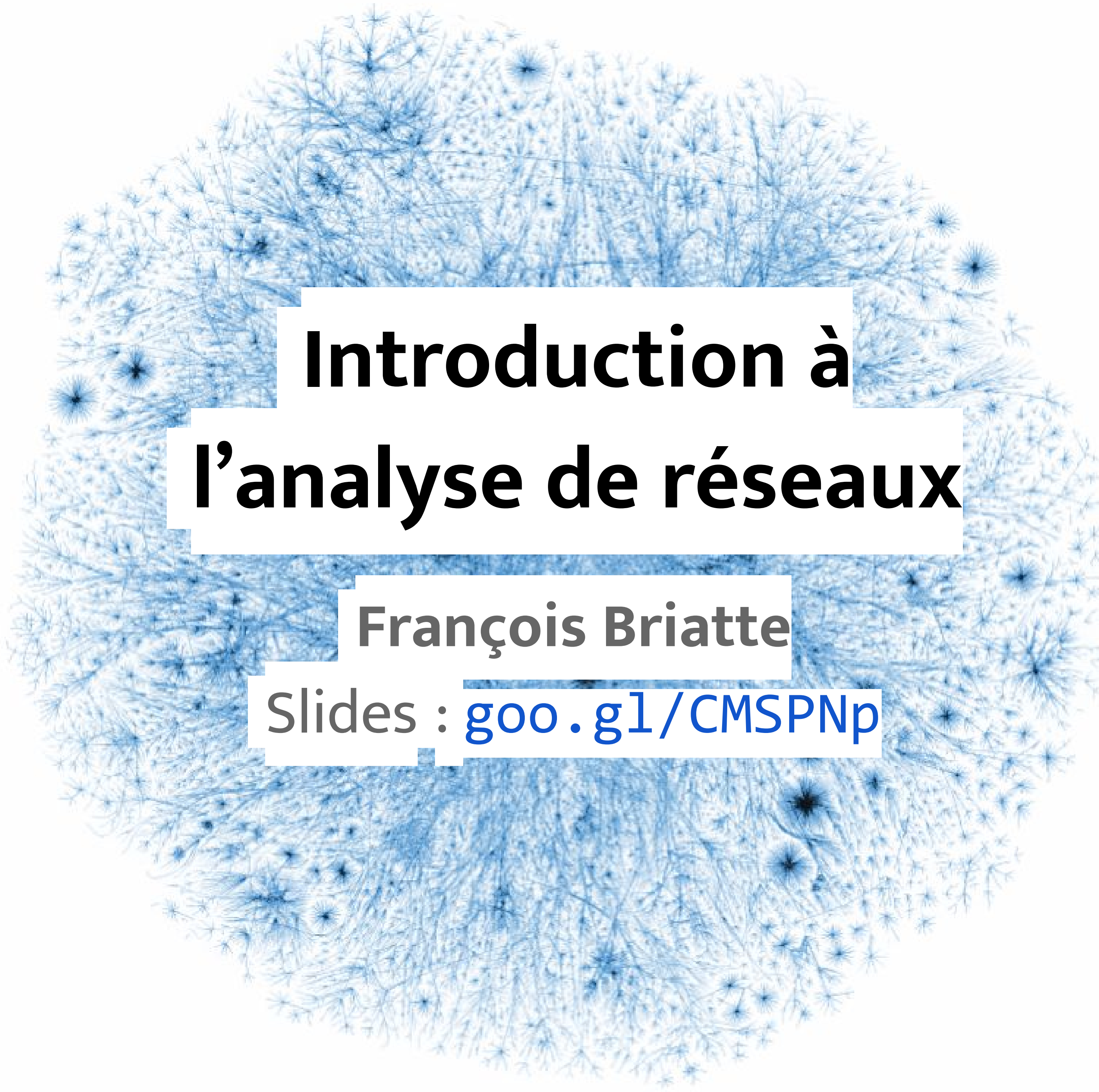
Introduction à l’analyse de réseaux



François Briatte

Slides : `goo.gl/CMSPNp`

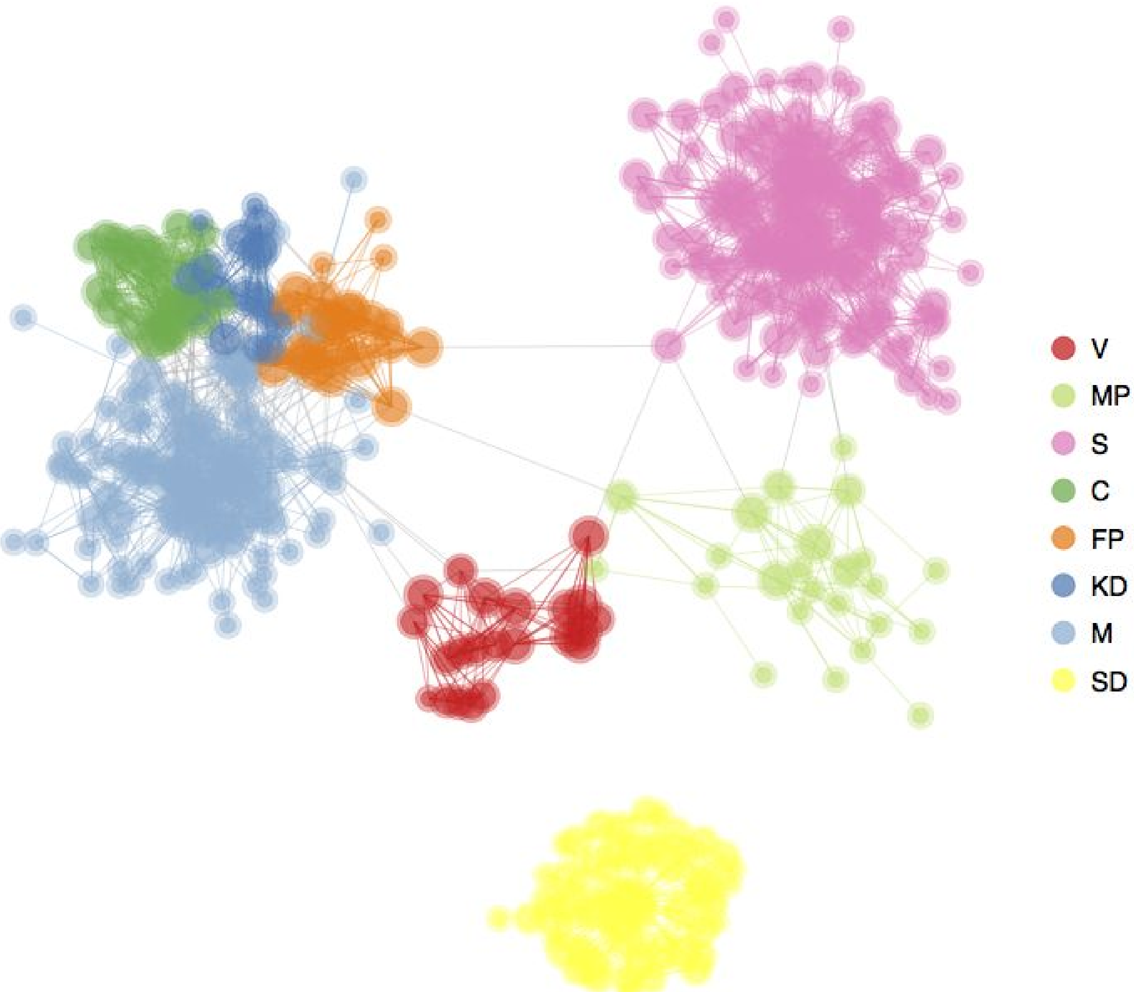
Ressources essentielles

- Un peu de **mathématiques** appliquées
  (tout part de là... et tout y revient)
- Un peu de **code** : **R** et le package `igraph`
  (qui existe aussi pour C et Python)
- Des tonnes de **données réelles**
  (pour éprouver les mesures et les modèles)

Exemple réel

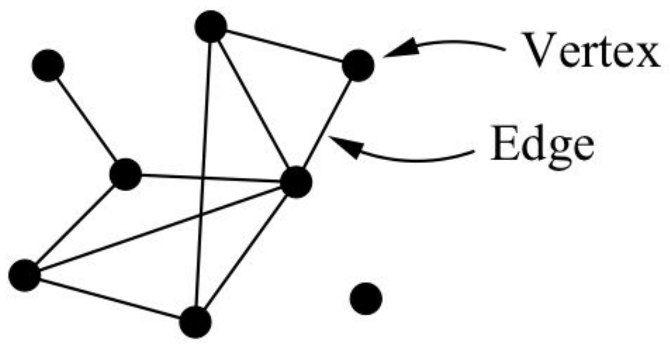
Cosignatures des propositions de loi au parlement suédois depuis 2014

Réseau unipartite, dirigé et pondéré, réalisé avec R et les packages `network` et `ggplot2`



Définition simplifiée

- Un réseau est un ensemble de **points** reliés par un ensemble de **lignes**
- La terminologie varie, mais la notation de base est toujours $G = (V, E)$ avec $V > 0$
- La définition exacte des « points » et des « lignes » appartient à l'analyste



Représentations numériques

- ***Adjacency matrix*** **:** matrices carrées de 0s et de 1s (ou autres valeurs) désignant l'absence, la présence (et possiblement la valeur) des liens entre les points du réseau

  (attention, adjacence ≠ incidence)

- ***Edge lists*** **:** séquence des liens existant dans le réseau (A, B), (B, C), (A, D), (C, D), (B, A), ...

Ce qu'il faut pouvoir établir

- Y a-t-il plusieurs **types de points** ?

  Réseaux **unipartites** (*one-mode*)

  Réseaux **bipartites** (*two-mode*) … ou k-partites avec $k > 1$

- Les liens entre les points ont-ils une **direction** ?

  Réseaux **dirigés** et **non dirigés**

- Les liens entre les points ont-ils des **valeurs** ?

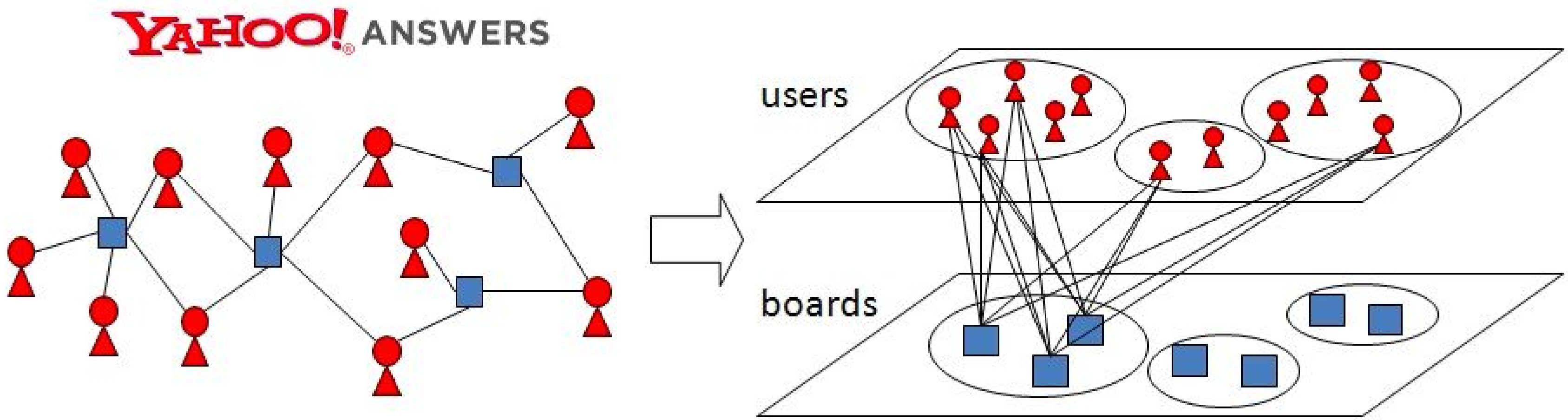
  Réseaux **binaires** et **pondérés**

  Réseaux **non signés** (+) et **signés** (±)

Ce qu'il faut aussi pouvoir déterminer

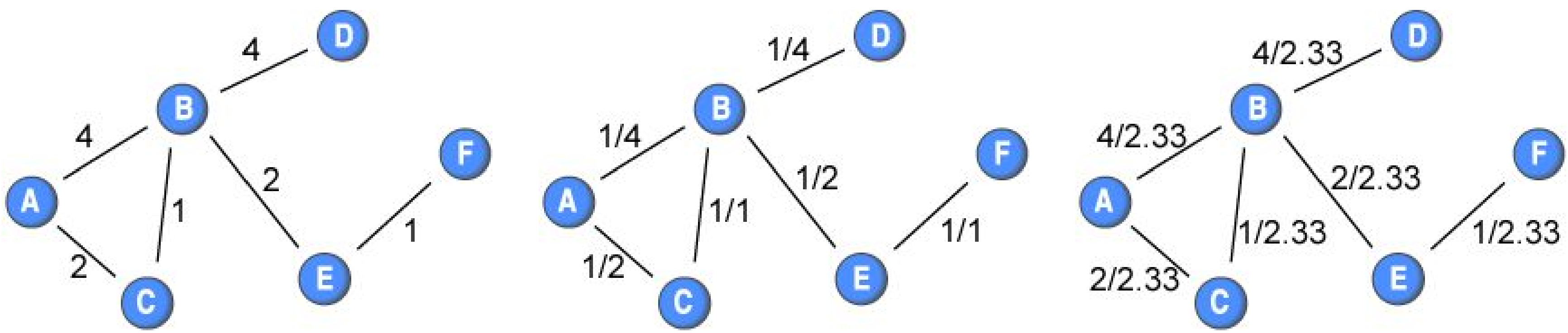
Dans un réseau bipartite, les membres d'un même mode peuvent-ils aussi créer **des liens entre eux** ?

Le cas échéant, tout se complique… !



Ce qu'il faut pouvoir imaginer

La pondération des liens d'un réseau n'a généralement rien d'évident...



... et une bonne pratique consiste à « tester » différentes manières de pondérer les liens.

Exemple réel

The equation that we implement is given in Gross, Kirkland and Shalizi (2012, eqn. 1, p. 8). Letting $c_{j(k)}$ denote the number of cosponsors on MP j's k^{th} bill, the strength of the tie between MP j and every cosponsor i on the bill is first weighted to $1/c_{jk}$, in order to downplay the influence of bills that are cosponsored by large numbers of MPs. At that stage, each first author is connected to each of his or her cosponsors by the sum of these weights, defined as

$$W_{ij} = \sum_{k=1}^{n_j} \frac{Y_{ij(k)}}{c_{j(k)}} \tag{E1}$$

(pondération par simple inversion)

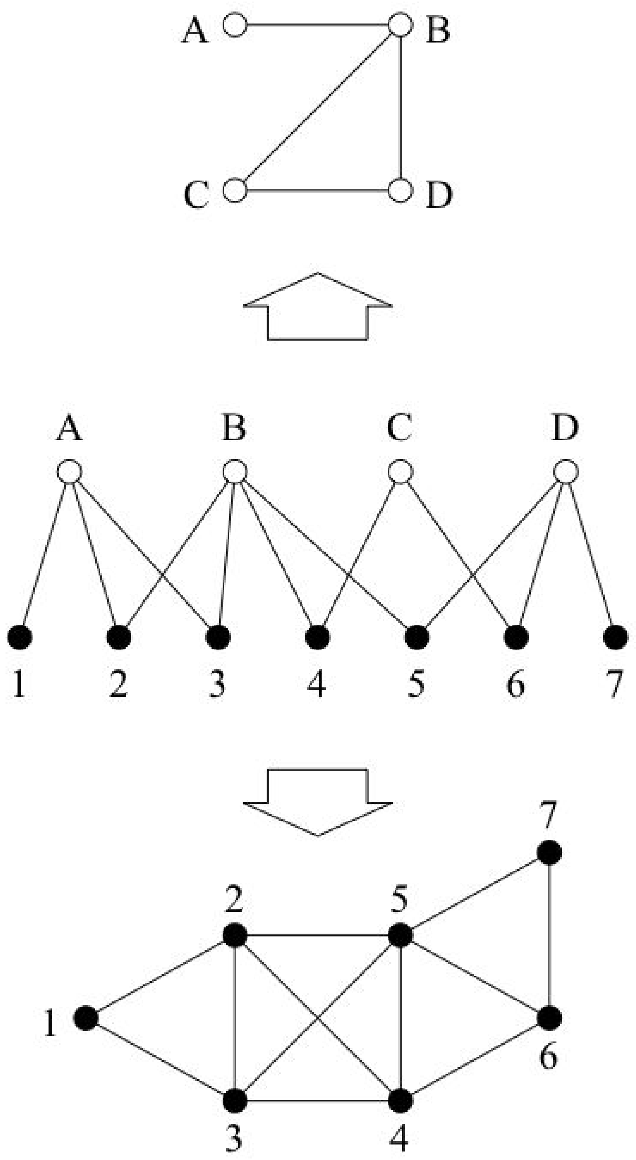
Exemple réel

The next step consists in dividing these weights, which Fowler (2006a, p. 468) calls the "weighted quantity of bills cosponsored", by the maximum value that they reach when MP i is a cosponsor on every k^{th} bill by MP j. The resulting weights, which Gross, Kirkland and Shalizi (2012, p. 8) call the "weighted propensity to cosponsor" and which we refer to as 'Gross-Shalizi weights' in reference to an earlier version of the manuscript by the two authors, are defined as

$$WPC_{ij} = \frac{\sum_{k=1}^{n_j} \frac{Y_{ij(k)}}{c_{j(k)}}}{\sum_{k=1}^{n_j} \frac{1}{c_{j(k)}}} \tag{E2}$$

(pondération « probabiliste »)

Principe des projections

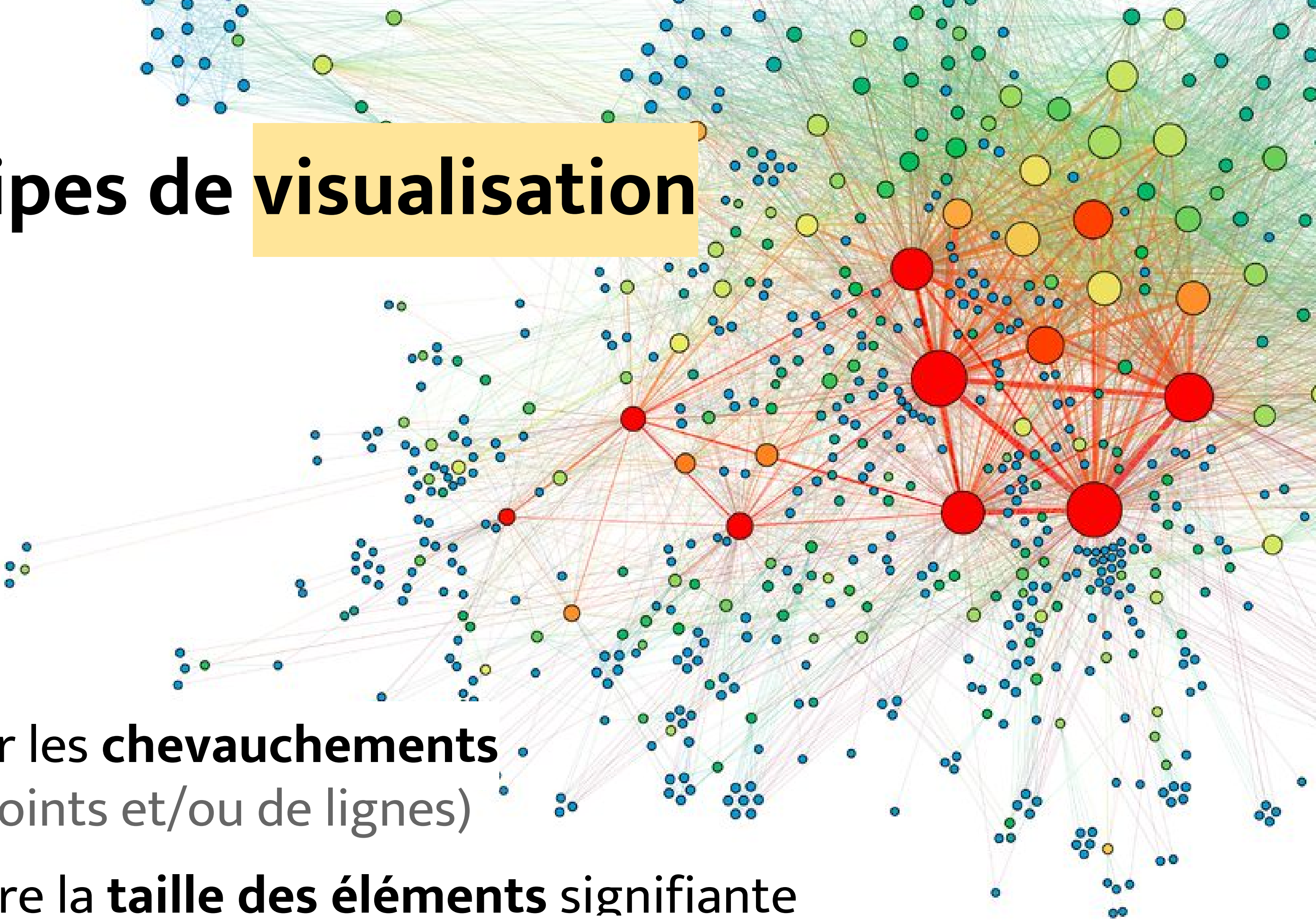


Un réseau bipartite peut être projeté sous la forme d'un réseau unipartite.

L'opération **détruit de l'information** mais peut notamment faciliter la **lecture graphique** des relations.

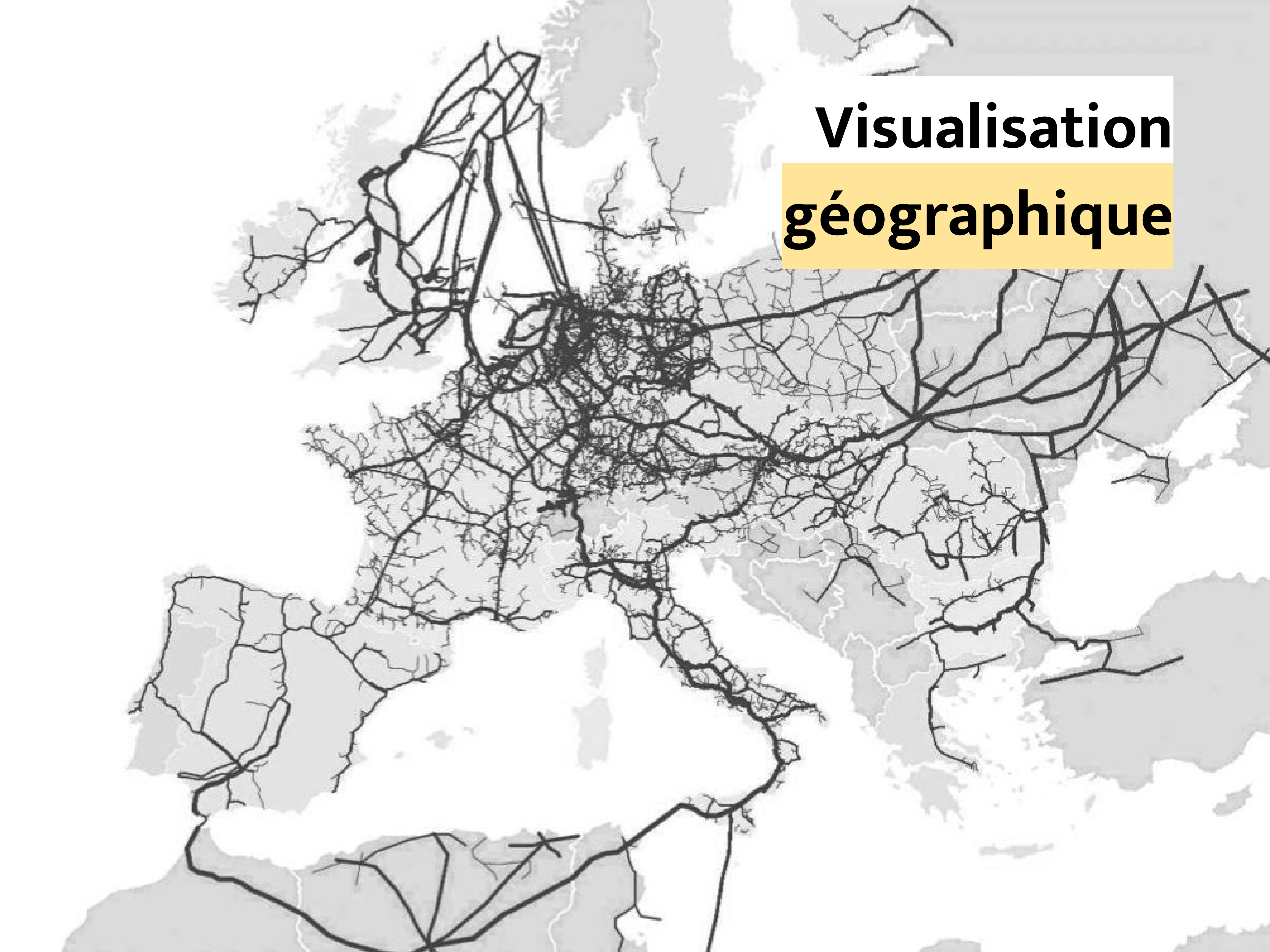
Figure 6.5: The two one-mode projections of a bipartite network. The central portion of this figure shows a bipartite network with four vertices of one type (open circles labeled A to D) and seven of another (filled circles, 1 to 7). At the top and bottom we show the one-mode projections of the network onto the two sets of vertices.

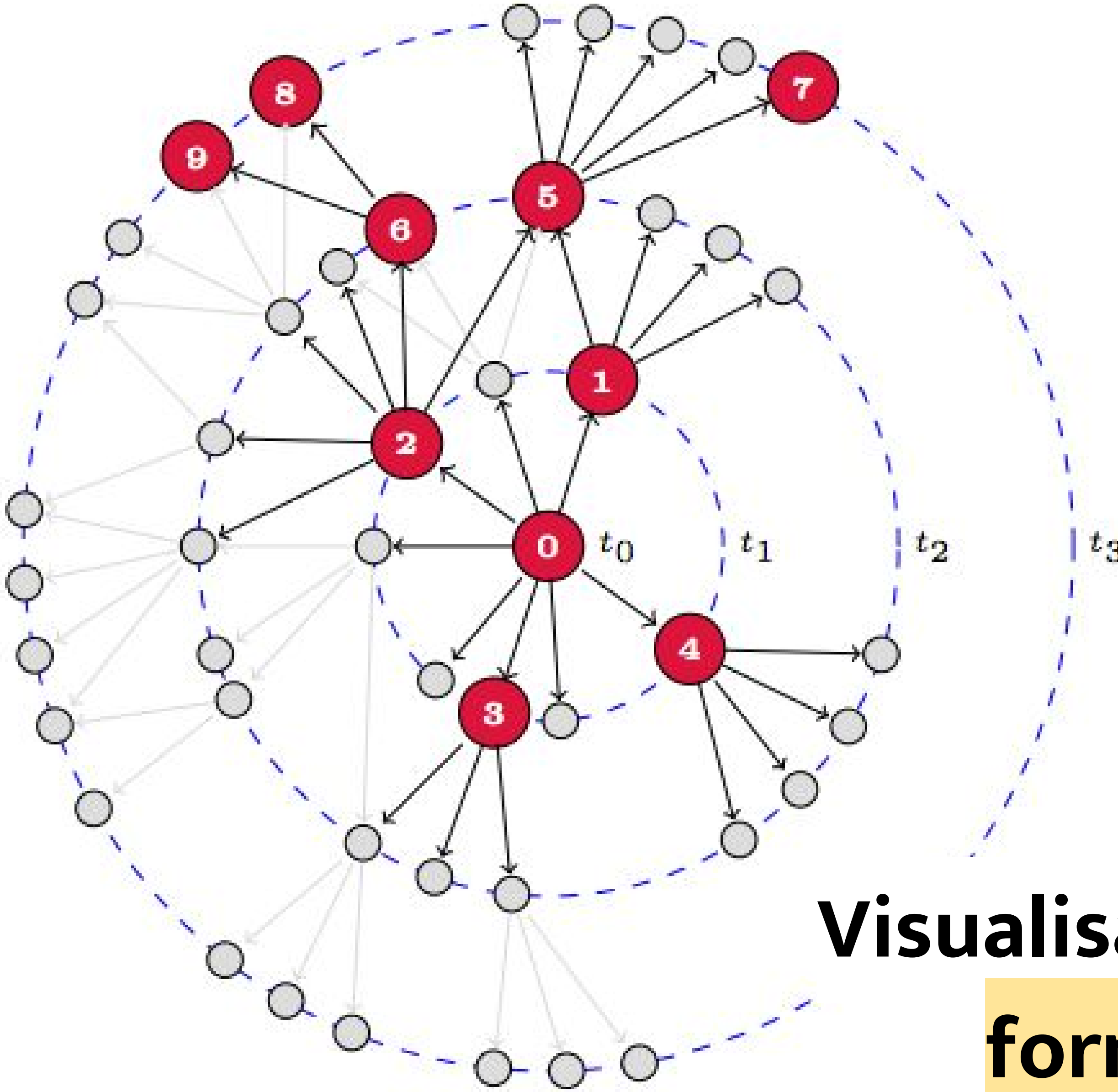
Principes de visualisation



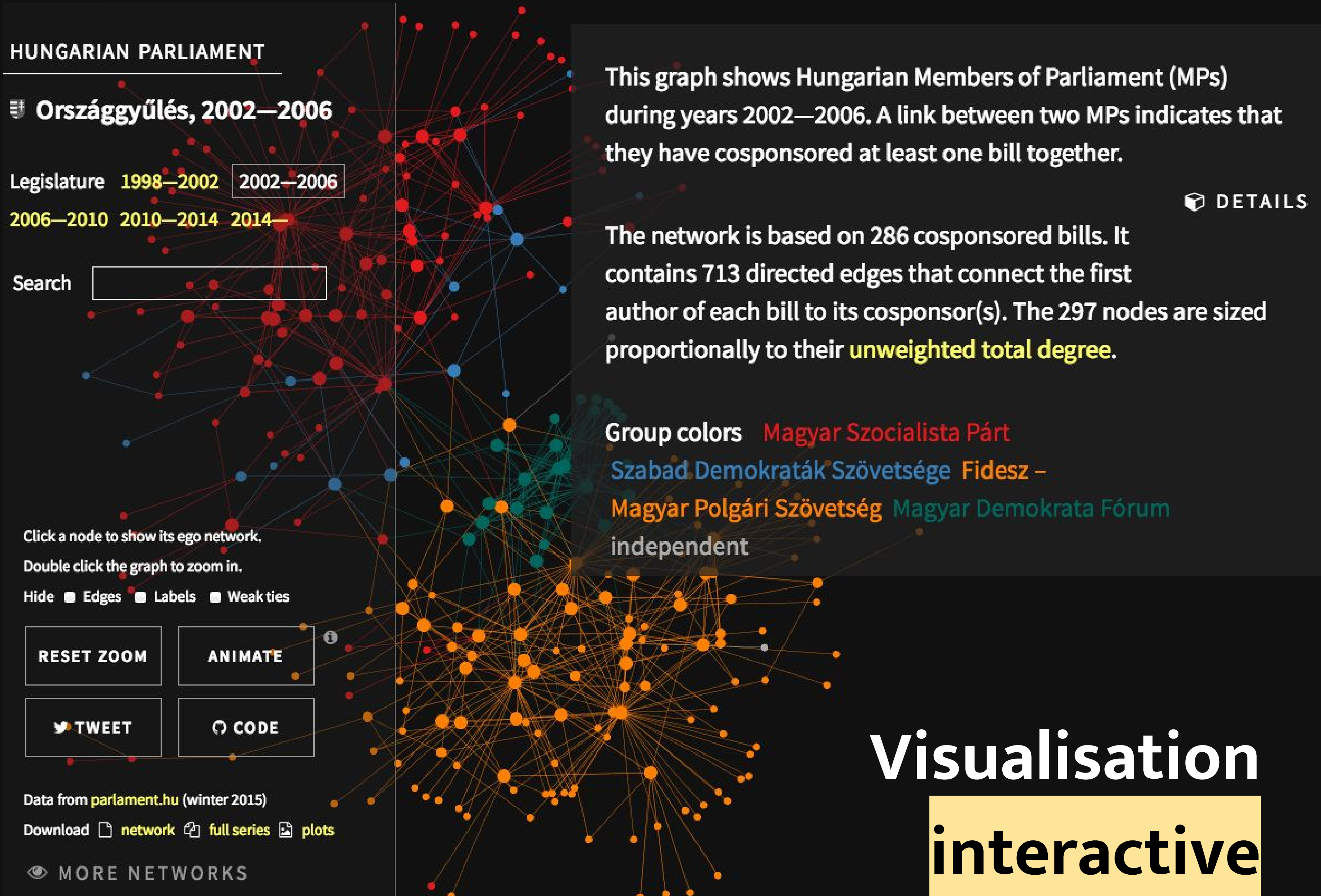
- Éviter les **chevauchements** (de points et/ou de lignes)
- Rendre la **taille des éléments** signifiante (taille des points, épaisseur des liens)
- Rendre la **distance des points** signifiante (proximité/distance)

Visualisation géographique





Visualisation formelle



Visualisation

interactive

Mesures applicables

- Mesures pour caractériser **le réseau** lui-même

  Nombre de points, nombre de liens

  Densité : nombre de liens / nombre total de liens possibles

- Mesures pour caractériser **les points**

  Degré : nombre de liens (si dirigé : entrants/sortants)

  Le degré est une des mesures de la **centralité** des points

Pourquoi parler de centralité ?

- L'idée est qu'un point peut être **influent**…

  « **Points de passage obligés** » d'un réseau informationnel

  Algorithme PageRank : la « force des liens rares »

- … y compris dans un **sens négatif**

  Points de vulnérabilité dans un réseau électrique

  Bottlenecks : goulots d'étranglement ou de ralentissement

Là encore, savoir choisir et tester…

	with constant term	without constant term
divide by out-degree	$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ PageRank	$\mathbf{x} = \mathbf{A}\mathbf{D}^{-1}\mathbf{x}$ degree centrality
no division	$\mathbf{x} = (\mathbf{I} - \alpha\mathbf{A})^{-1} \cdot \mathbf{1}$ Katz centrality	$\mathbf{x} = \kappa_1^{-1}\mathbf{A}\mathbf{x}$ eigenvector centrality

Table 7.1: Four centrality measures. The four matrix-based centrality measures discussed in the text are distinguished by whether or not they include an additive constant term in their definition and whether they are normalized by dividing by the degrees of neighboring vertices. Note that the diagonal matrix $\mathbf{D}$, which normally has elements $D_{ii} = k_i$, must be defined slightly differently for PageRank, as $D_{ii} = \max(1, k_i)$—see Eq. (7.15) and the following discussion. Each of the measures can be applied to directed networks as well as undirected ones, although only three of the four are commonly used in this way. (The measure that appears in the top right corner of the table is equivalent to degree centrality in the undirected case but takes more complicated values in the directed case and is not widely used.)

D’autres mesures à connaître

- **Distance la plus courte** (*shortest path*) entre deux points

  (généralement assez facile à calculer)

- **Homophilie** : propensité des liens à renvoyer à des attributs partagés par les points

  (souvent beaucoup moins simple à calculer)

D’autres méthodes à connaître

- La mesure d’une probabilité suppose de disposer d’une « hypothèse nulle » ; du coup, on étudie aussi les **réseaux aléatoires**
- Tester cette « hypothèse nulle » peut signifier : minimiser un ou des paramètre(s) ; du coup, on utilise des **algorithmes d’optimisation**

Exemple

La **modularité** d’un réseau compare la création de liens entre les points d’une même communauté à celle de liens entre communautés différentes :

$$\text{Mod} = \frac{1}{2m}\sum_{ij}[A_{ij} - \frac{k_i k_j}{2m}] * \delta(c_i, c_j), \qquad (1)$$

where m is the total number of connections in the network, A_{ij} is the connection between actors i and j, k_i is the total number of connections actor i has (the degree of actor i), and $\delta(c_i, c_j)$ is the Kroneker delta for communities of actors i and j.[4] Thus, modularity examines all the actors in the same community $\delta(c_i, c_j)$, adds up their connections with one another (A_{ij}), and subtracts the connections we might expect those actors to have in the network generally ($k_i k_j / 2m$). Finally, the sum of all these components is normalized by the total number of connections in the network ($1/2m$).

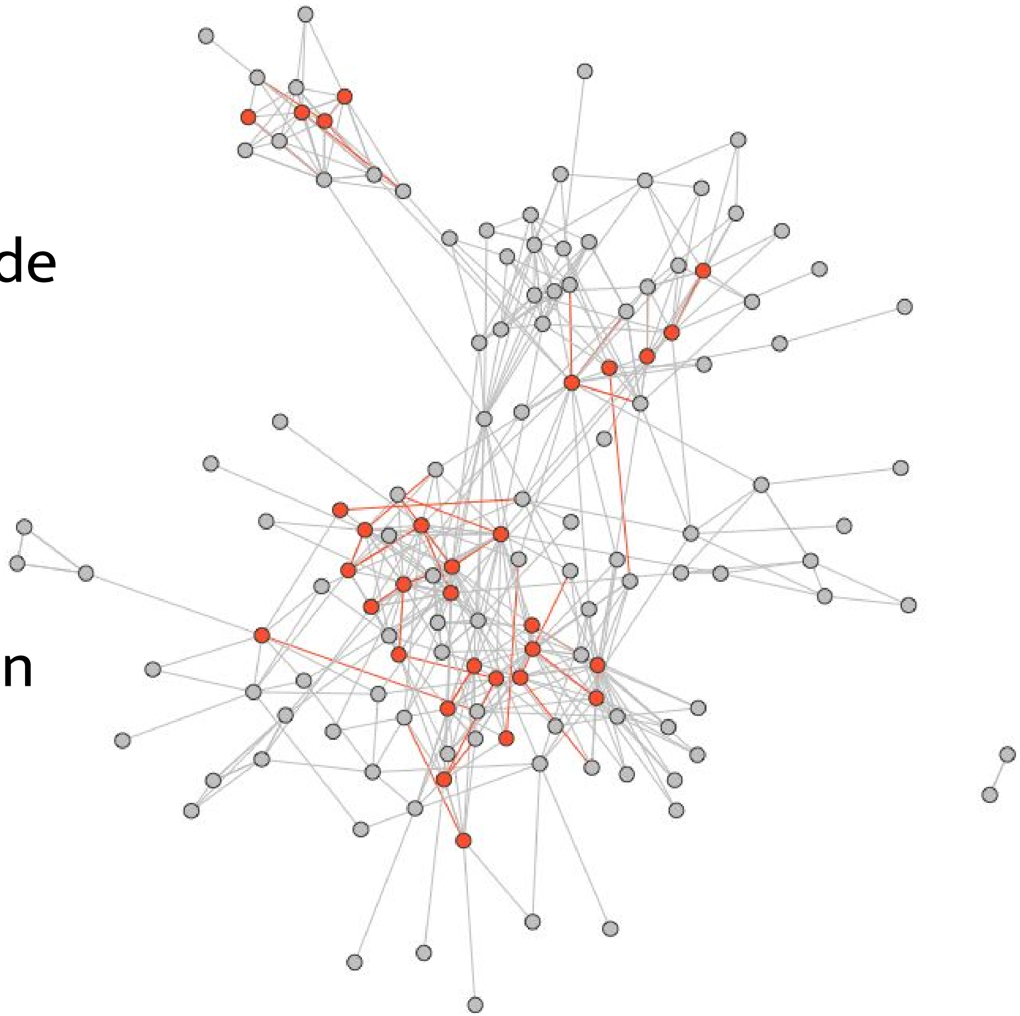
Exemple

La **détection de communautés** partitionne les points d'un réseau en cherchant les communautés qui minimisent la modularité du réseau :

Et encore d’autres méthodes…

La **percolation** d’un réseau consiste à détruire progressivement ses liens, de différentes manières.

La recherche de « l’épine dorsale » – ***backbone*** – d’un réseau consiste à chercher son composant le plus « résilient » (robuste).



Pour s'y retrouver

- Commencer par les **questions les plus simples**

  De quoi est ***vraiment*** constitué le réseau ?

  Quelles sont les **interactions** entre les points ?

- Déterminer le **point d'arrivée** de l'analyse

  Des **descripteurs** pour les points (ou les liens) du réseau

  Des **prédicteurs** de l'état du réseau

Pourquoi est-ce important

- À **données bruitées**, mesures bruitées

  Le réseau contient-il bien **tous les liens** formés ?

  Quelle est la proportion de **liens accidentels** ?

- **Mesures** ≠ **modèles**

  Mesurer les « **conséquences** » des liens (ex. centralité)

  ≠ Comprendre **le mécanisme** de génération des liens

Points de discussion

- Bien étudier la **morphologie des données**

  Combien de **modes/niveaux** les données reflètent-elles ?

  À quel niveau (pour quel mode) faut-il étudier les liens ?

- Bien comprendre les ***trade-offs*** de l'analyse

  La visualisation contient souvent un **composant aléatoire**

  La visualisation peut exiger de **perdre de l'information**

Points de discussion

- Les **attributs** du réseau sont essentiels

  Quels sont les attributs essentiels **des points** ?

  Quels sont les attributs essentiels **des liens** ?

- La **visualisation** d'un réseau est soit…

  … une fin en soi, auquel cas **l'image arrête la pensée**

  … un point de départ pour **imaginer un modèle**

Liens

- Awesome Network Analysis

  Une liste avec plein, *plein* de ressources utiles, avec notamment plusieurs **ressources en langue française**.

- Network Analysis and Visualization with R and igraph

  Le meilleur tutoriel de **visualisation de réseaux**, par Katherine Ognyanova.

Références

- **Newman**
  Networks. An Introduction (2010)
- **Samatova et al.** (eds)
  Practical Graph Mining with R (2014)
- **Kolaczyk, Csárdi**
  Statistical Analysis of Network Data with R (2014)

Exemples détaillés

- **Hypothesosphère :** le réseau des blogs d’Hypothèses en 2015 politbistro.hypotheses.org/2737
- **Les revues de Cairn.info :** réseau et mesures politbistro.hypotheses.org/tag/cairn
- **Notices Calenda :** invités et organisateurs (non publié, travail en cours… !)

Merci pour votre attention

francois.briatte@sciencespo.fr