Groupes et communautés dans les flots de liens : des données aux algorithmes

Noé Gaumont
Encadrants: Clémence Magnien et Matthieu Latapy

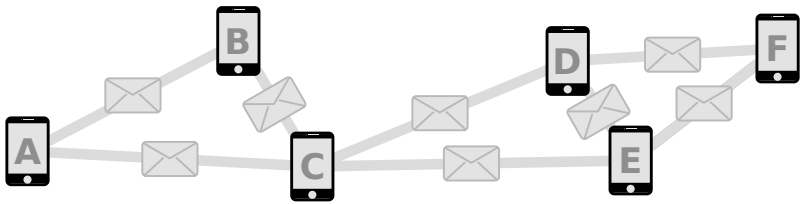
LIP6 - CNRS & UPMC , Sorbonne Universités.

11 Octobre 2016

Réseaux complexes

Réseau complexe : ensemble d'éléments en interaction.

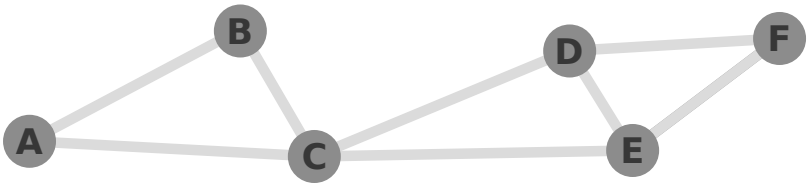
réseaux	éléments	interactions
cerveau	régions du cerveau	influx nerveux
trafic IP	ordinateurs	paquets IP
télécommunication	téléphones	appels/SMS



Réseaux complexes

Réseau complexe : ensemble d'éléments en interaction.

réseaux	éléments	interactions
cerveau	régions du cerveau	influx nerveux
trafic IP	ordinateurs	paquets IP
télécommunication	téléphones	appels/SMS

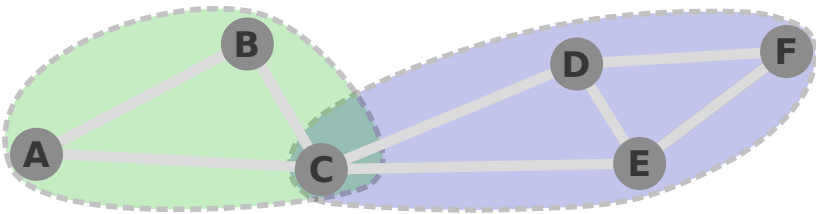


Théorie des graphes

Réseaux complexes

Réseau complexe : ensemble d'éléments en interaction.

réseaux	éléments	interactions
cerveau	régions du cerveau	influx nerveux
trafic IP	ordinateurs	paquets IP
télécommunication	téléphones	appels/SMS



Théorie des graphes
Détection de communautés

Réseaux complexes

Réseau complexe : ensemble d'éléments en interaction.

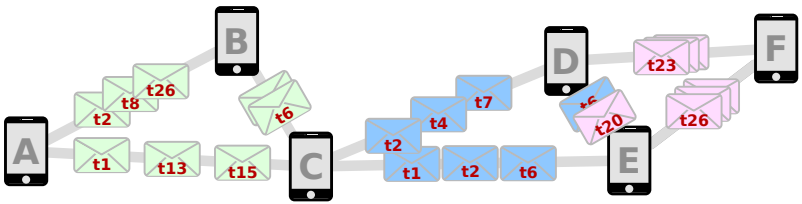
réseaux	éléments	interactions
cerveau	régions du cerveau	influx nerveux
trafic IP	ordinateurs	paquets IP
télécommunication	téléphones	appels/SMS

Modèle prenant en compte la temporalité ?
Qu'est-ce qu'une communauté ?

Réseaux complexes

Réseau complexe : ensemble d'éléments en interaction.

réseaux	éléments	interactions
cerveau	régions du cerveau	influx nerveux
trafic IP	ordinateurs	paquets IP
télécommunication	téléphones	appels/SMS



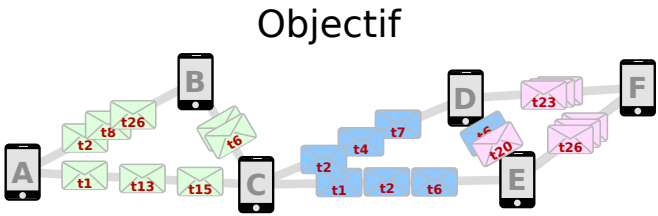
Formalisme de flots de liens
Partitions des interactions

Contributions

Réseaux statiques

Évaluation de partitions de liens

- Modèle nul
- Fonction de qualité
- Tests
- Algorithme d'optimisation



Réseaux dynamiques

Formalisme

- Flots de liens
- Représentation
- Implémentaion

Description de données réelles

- Densité externe
- Flot quotient
- Détection de structure automatique

Détection de groupes pertinents

- Densité
- Pertinence
- Références
- Applications

Vers des partitions de flots de liens

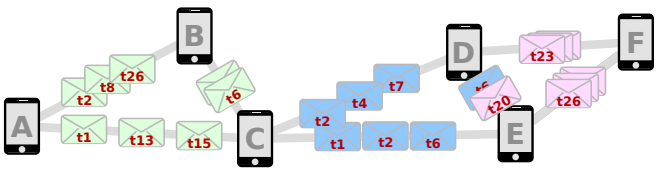
- Générations
- Étude de méthodes statiques
- Première fonction de qualité

Contributions

Réseaux statiques

Évaluation de partitions de liens
- Modèle nul
- Fonction de qualité
- Tests
- Algorithme d'optimisation

Objectif



Réseaux dynamiques

Formalisme
- Flots de liens
- Représentation
- Implémentaion

Description de données réelles
- Densité externe
- Flot quotient
- Détection de structure automatique

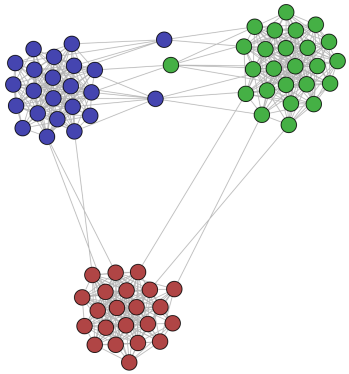
Détection de groupes pertinents
- Densité
- Pertinence
- Références
- Applications

Vers des partitions de flots de liens
- Générations
- Étude de méthodes statiques
- Première fonction de qualité

Évaluation d'une partition de liens d'un graphe

Communautés de nœuds
Communautés de liens
Définition de *Expected Nodes*
Tests avec le générateur LF
Conclusions

Communautés de nœuds



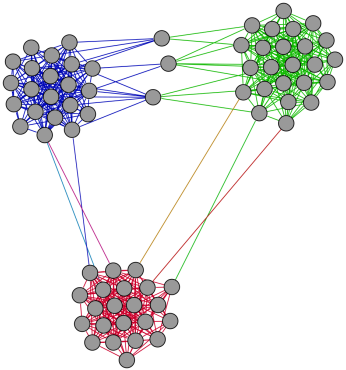
Graphe : réseau téléphonique.
Communautés : groupes de personnes.

Entrée :
Un graphe, $G = (V, E)$.
Une partition de **nœuds**.

Sortie :
Évaluation de la partition en tant que structure communautaire.

La modularité pour évaluer une partition de nœuds.

Communautés de liens

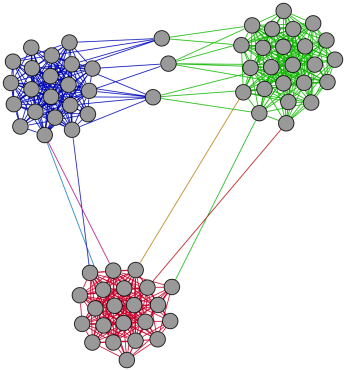


Graphe : réseau téléphonique.
Communautés : discussions entre personnes.

Entrée :
Un graphe, $G = (V, E)$.
Une partition de **liens**.

Sortie :
Évaluation de la partition en tant que structure communautaire.

Communautés de liens



Graphe : réseau téléphonique.
Communautés : discussions entre personnes.

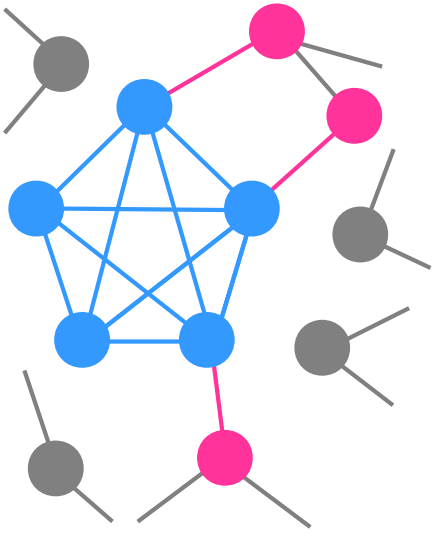
Entrée :
Un graphe, $G = (V, E)$.
Une partition de **liens**.

Sortie :
Évaluation de la partition en tant que structure communautaire.

Proposition de **Expected Nodes** pour évaluer une partition de liens.

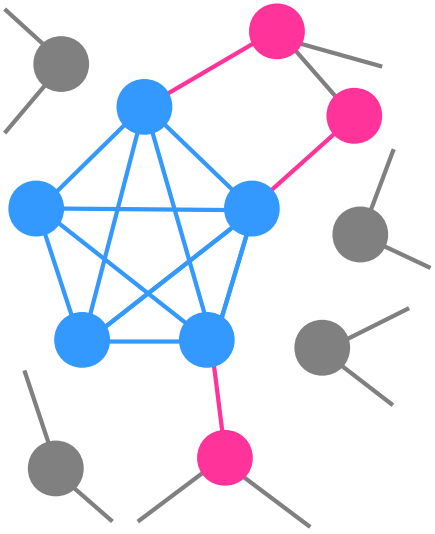
Intuition sur la fonction de qualité

En quoi le groupe de liens bleus est-il intéressant ?



Intuition sur la fonction de qualité

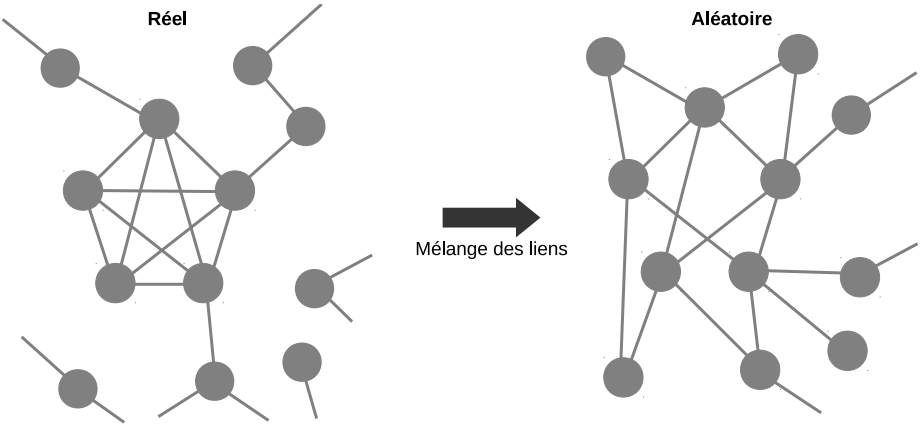
En quoi le groupe de liens bleus est-il intéressant ?



Les liens bleus sont très denses.
Les liens adjacents en roses sont peu denses.

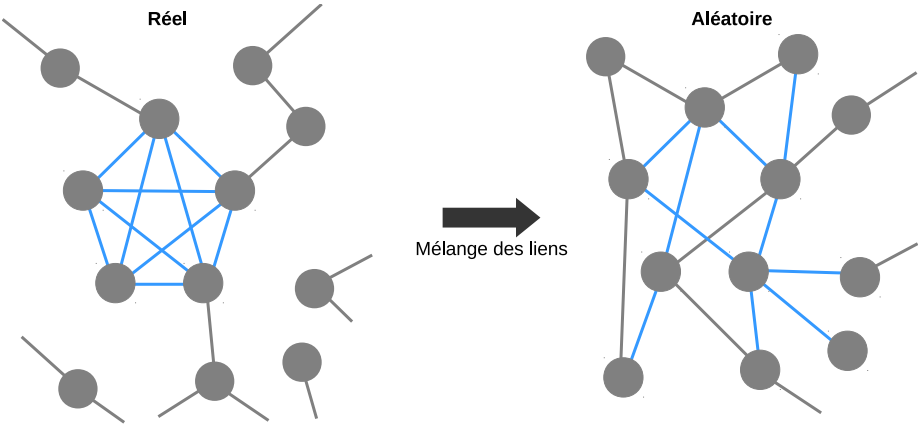
Qualité interne

Comparer le nombre de nœuds internes observé et celui attendu :



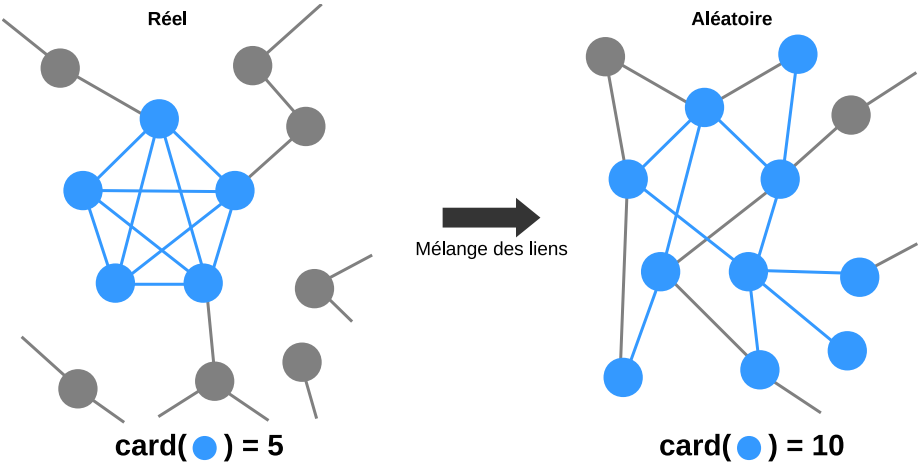
Qualité interne

Comparer le nombre de nœuds internes observé et celui attendu :



Qualité interne

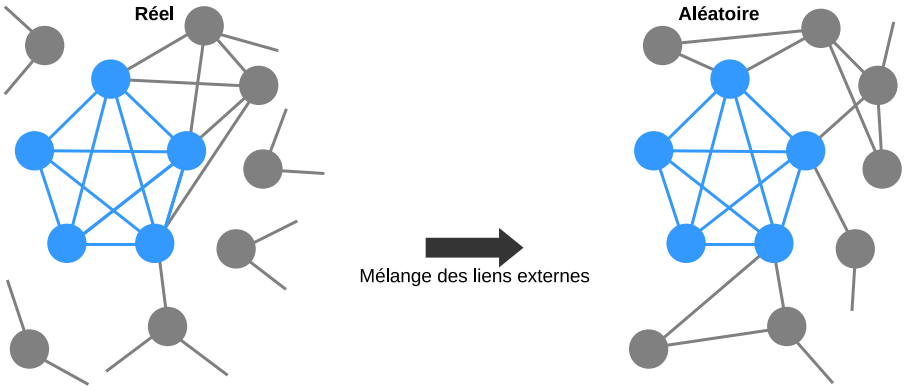
Comparer le nombre de nœuds internes observé et celui attendu :



$$Q_{in}(L) = \frac{\mathbb{E}[V(L)] - |V(L)|}{\mathbb{E}[V(L)]} = \frac{10-5}{10} = 0.5$$

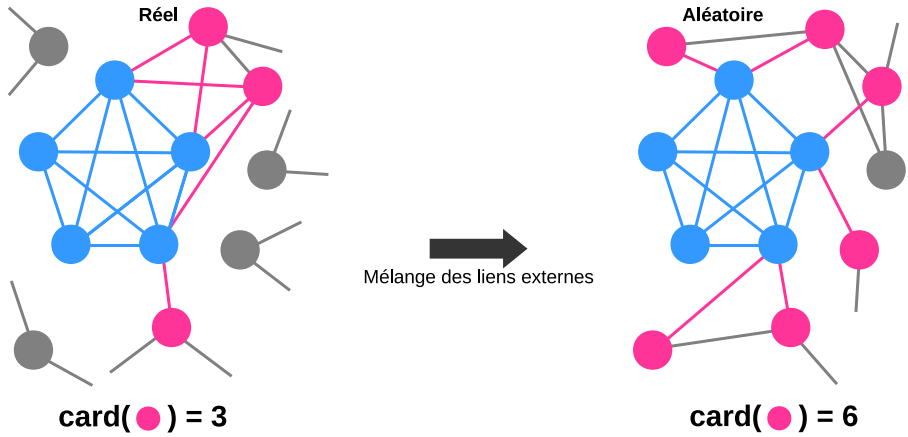
Qualité externe

Comparer le nombre de nœuds externes observé et celui attendu :



Qualité externe

Comparer le nombre de nœuds externes observé et celui attendu :

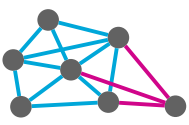


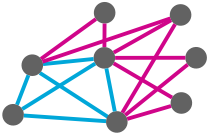
$$Q_{ext}(L) = \frac{|V_{ext}(L)| - \mathbb{E}[V_{ext}(L)]}{\mathbb{E}[V_{ext}(L)]} = \frac{3-6}{6} = -0.5$$

Combinaison des qualités internes et externes

Qualité d'un groupe de liens L :

$$Expected\ Nodes(L) = 2\frac{|L|Q_{in}(L) + |L_{ext}|Q_{ext}(L_{ext})}{|L| + |L_{ext}|}$$

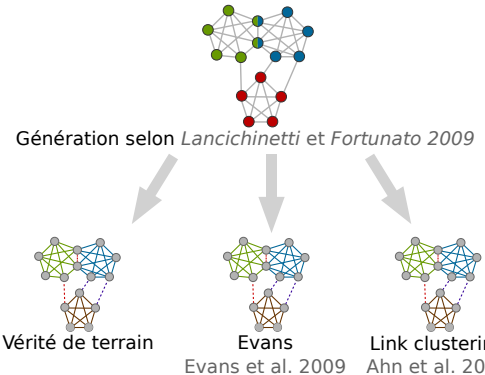




Expected Nodes(L) > *Expected Nodes(L)*

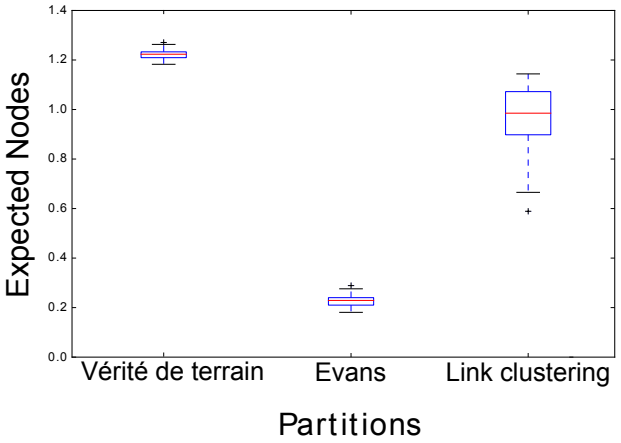
Méthodologie

Générations de graphes ayant une structure communautaire sur les **nœuds**.



Sur chaque graphe, évaluation par **Expected Nodes** des trois partitions.

Résultats



La vérité de terrain a une meilleure évaluation que les partitions *Evans* et *Link Clustering*.

Conclusions

En résumé

- Étude des communautés de liens au lieu de nœuds.
- Définition d'une nouvelle fonction de qualité : *Expected Nodes*.
- Sur nos tests, *Expected Nodes* met en avant la vérité de terrain.
- Les méthodes existantes n'ont pas cette caractéristique.

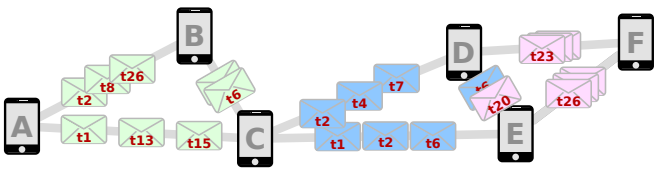
Contributions

Réseaux statiques

Évaluation de partitions de liens

- Modèle nul
- Fonction de qualité
- Tests
- Algorithme d'optimisation

Objectif



Réseaux dynamiques

Formalisme

- Flots de liens
- Représentation
- Implémentaion

Description de données réelles

- Densité externe
- Flot quotient
- Détection de structure automatique

Détection de groupes pertinents

- Densité
- Pertinence
- Références
- Applications

Vers des partitions de flots de liens

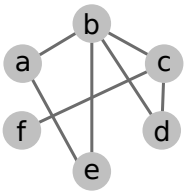
- Générations
- Étude de méthodes statiques
- Première fonction de qualité

Formalisme de flot de liens

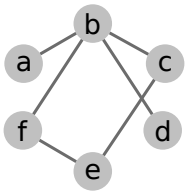
Séries de graphes
Graphes temporels
Flots de liens

Séries de graphes

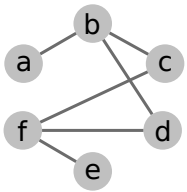
Série de graphes : ensemble de graphes statiques.



T=[0,3[



T=[3,6[

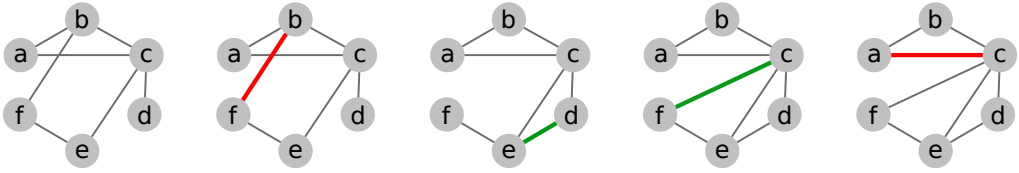


T=[6,9[

Nécessite de connaître une échelle de temps pertinente
Perte d'information

Graphes temporels

Graphe temporels : ensembles d'ajouts et de suppressions de liens



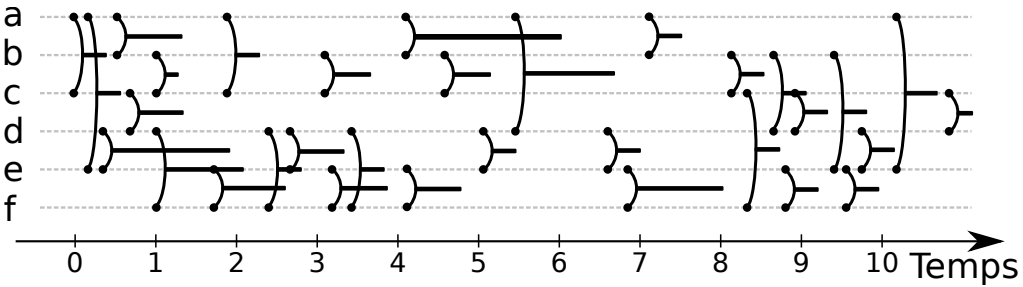
Pas de perte d'information temporelle

Structure de graphe à chaque instant

Flots de liens

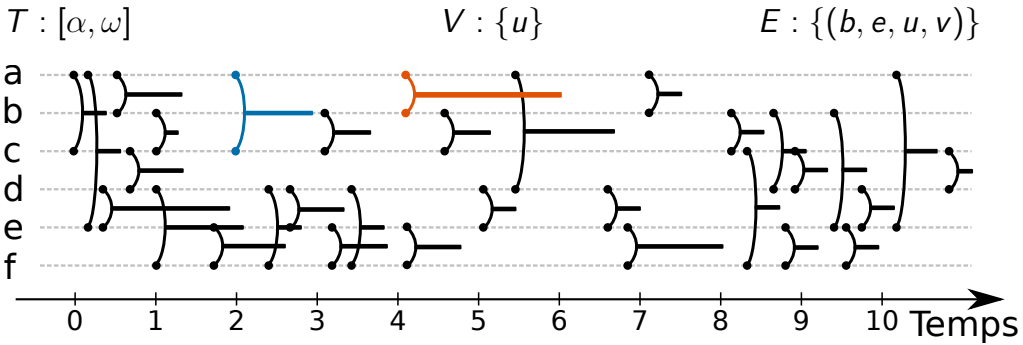
$L = (T, V, E)$

$T : [\alpha, \omega]$ $V : \{u\}$ $E : \{(b, e, u, v)\}$



Flots de liens

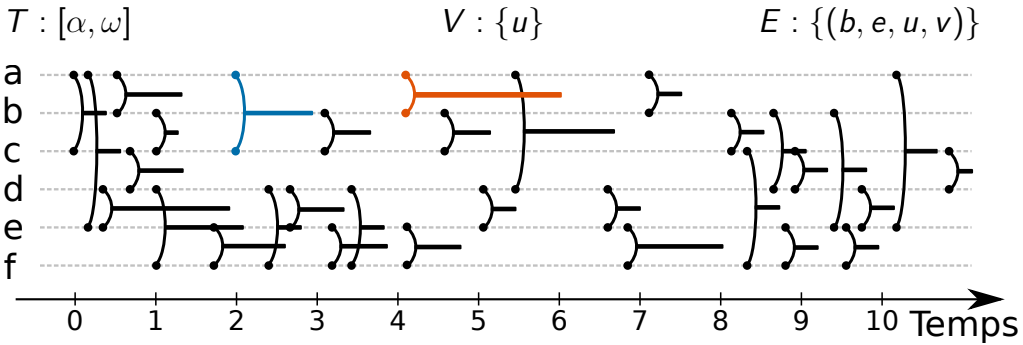
$L = (T, V, E)$



a et c interagissent sur [2,3] $\Rightarrow (2, 3, a, c) \in E$
a et b interagissent sur [4,6] $\Rightarrow (4, 6, a, b) \in E$

Flots de liens

$L = (T, V, E)$



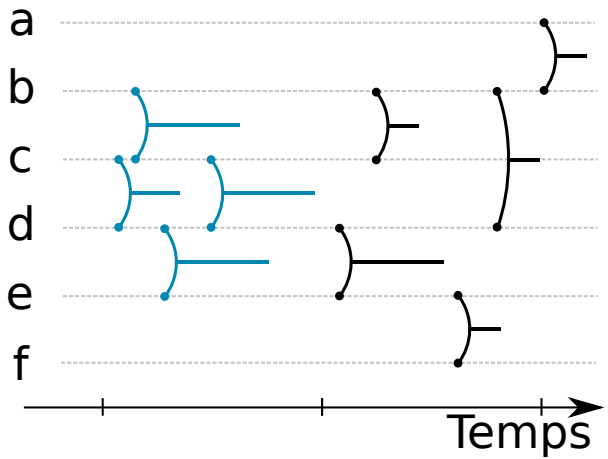
a et c interagissent sur [2,3] $\Rightarrow (2, 3, a, c) \in E$
a et b interagissent sur [4,6] $\Rightarrow (4, 6, a, b) \in E$

Pas de perte d'information temporelle
Définition de concepts adaptés aux flots de liens

Détection de groupes pertinents dans les flots de liens

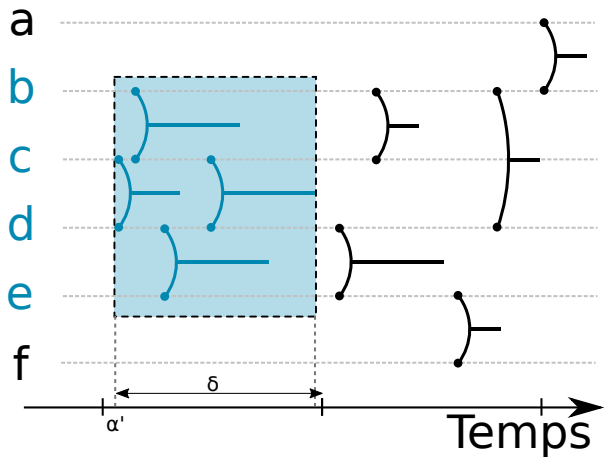
Définition de la pertinence
Détection de groupes pertinents
Jeux de données et applications
Conclusions

Densité d'un groupe de liens



Densité d'un groupe de liens

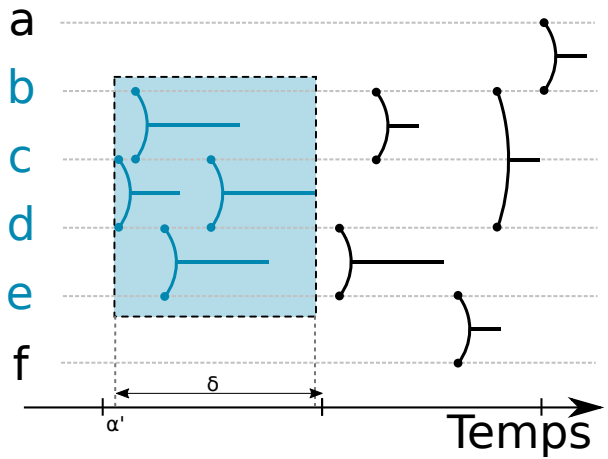
$V' = \{b, c, d, e\}$



$d(V', \alpha', \delta)$ = probabilité qu'il existe un lien entre 2 nœuds dans V' à un instant dans l'intervalle $[\alpha', \alpha' + \delta]$.

Densité d'un groupe de liens

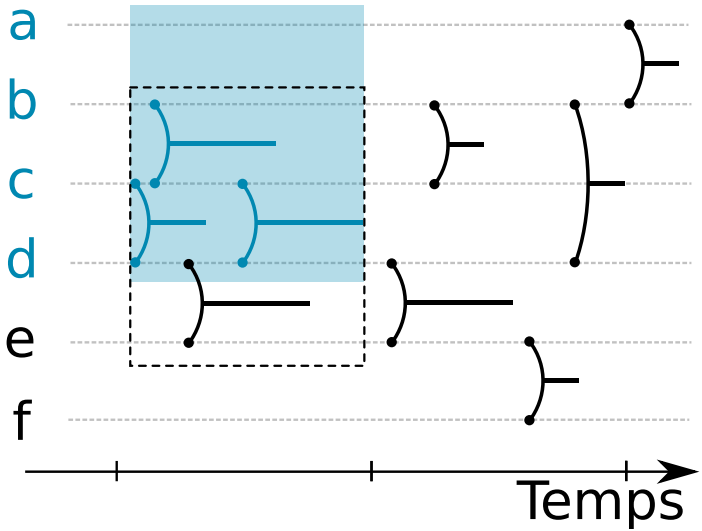
$V' = \{b, c, d, e\}$



Si $d = 0.13$, est-ce élevé ?

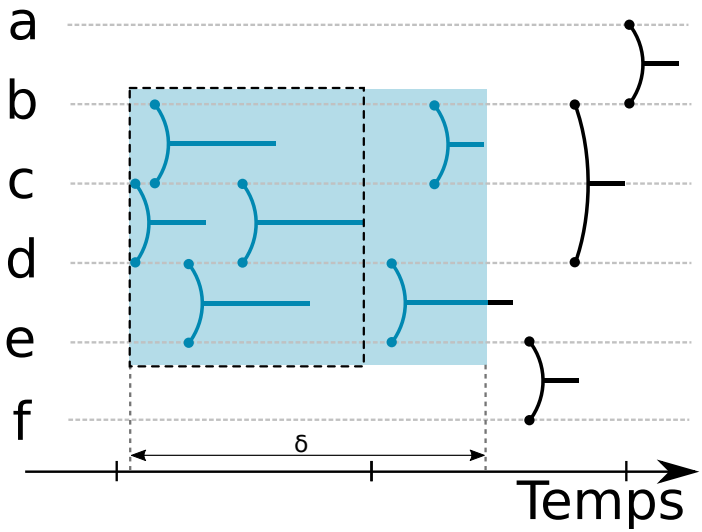
$d(V', \alpha', \delta)$ = probabilité qu'il existe un lien entre 2 nœuds dans V' à un instant dans l'intervalle $[\alpha', \alpha' + \delta]$.

Référence sur les nœuds



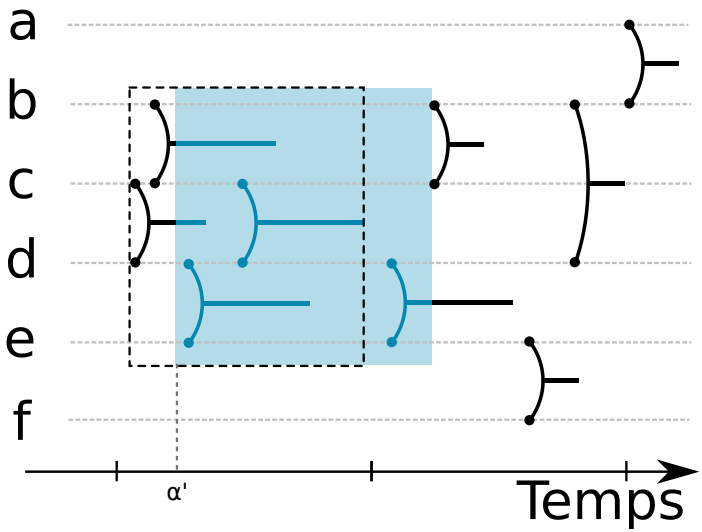
$d(\mathbf{X}, \alpha', \delta)$ tel que $|\mathbf{X} \cap V'| = |V'| - 1 = |\mathbf{X}| - 1$

Référence sur la durée



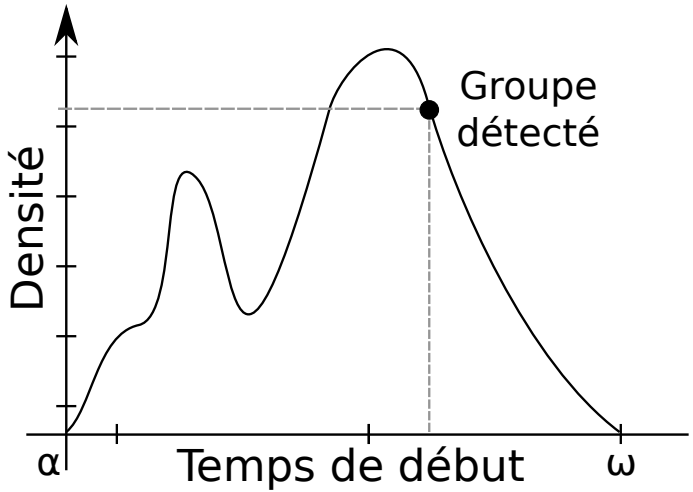
$d(V', \alpha', \mathbf{X})$ tel que $\mathbf{X} \in [0, \omega]$

Référence sur le temps de début

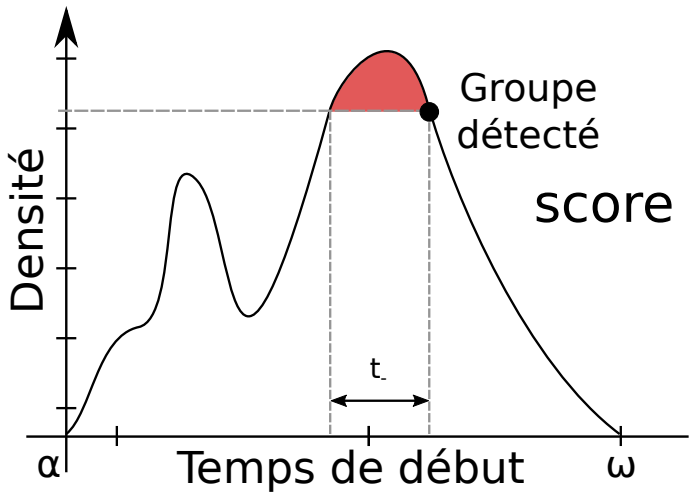


$d(V', \mathbf{X}, \delta)$ tel que $\mathbf{X} \in [\alpha, \omega]$

Évaluation d'un groupe



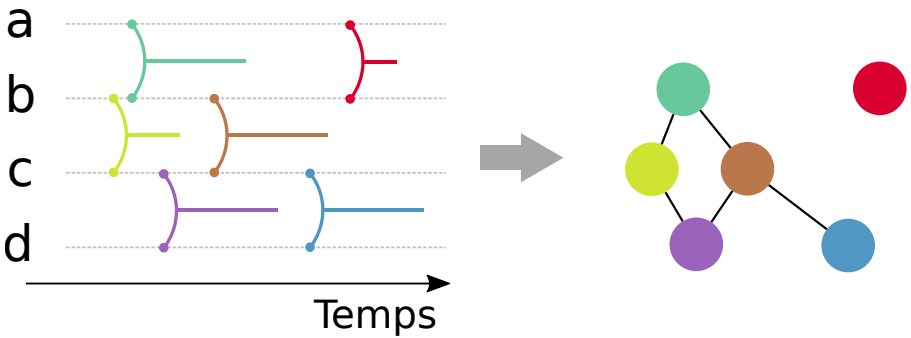
Évaluation d'un groupe



$$score = 1 - \frac{t_-}{T}$$

- $0 \leq score_{référence} \leq 1$
- Score élevé : le groupe est plus dense que la majorité des groupes pour la référence considérée.

Comment trouver des groupes potentiellement pertinents ?

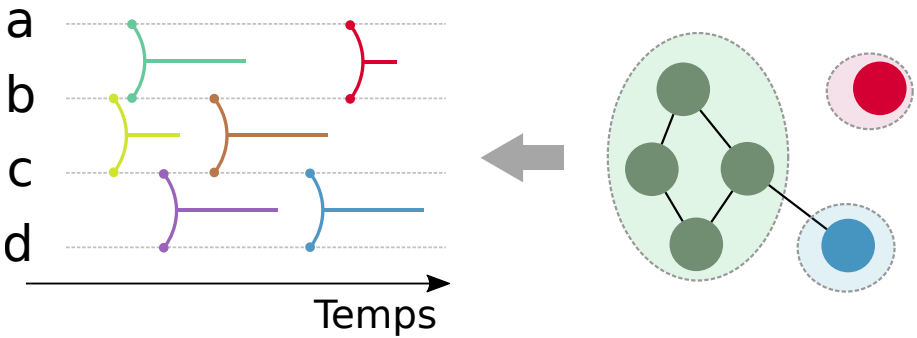


Flot de liens

Graphe statique

1. Créer le graphe statique.

Comment trouver des groupes potentiellement pertinents ?

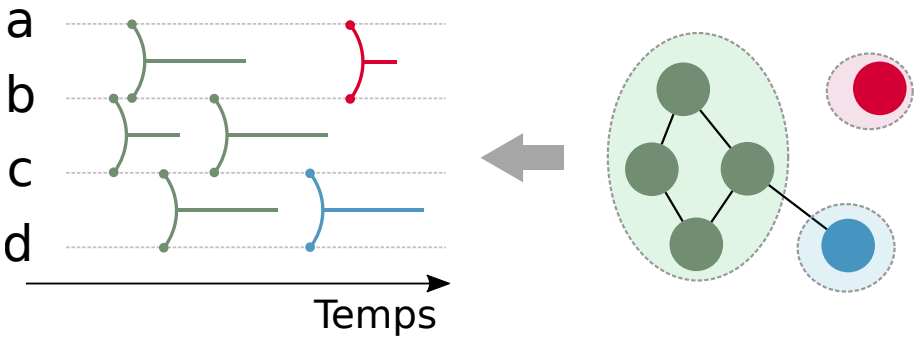


Flot de liens

Graphe statique

1. Créer le graphe statique.
2. Détection des communautés du graphe statique, via *Louvain*.

Comment trouver des groupes potentiellement pertinents ?

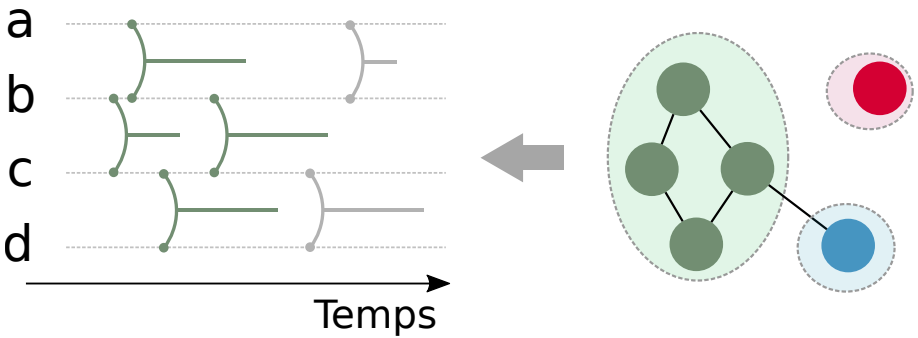


Flots de liens

Graphe statique

1. Créer le graphe statique.
2. Détection des communautés du graphe statique, via *Louvain*.
3. Transfert de la partition trouvée dans le flot de liens.

Comment trouver des groupes potentiellement pertinents ?



Flot de liens

Graphe statique

1. Créer le graphe statique.
2. Détection des communautés du graphe statique, via *Louvain*.
3. Transfert de la partition trouvée dans le flot de liens.
4. Application de la méthode d'évaluation pour garder certains groupes .

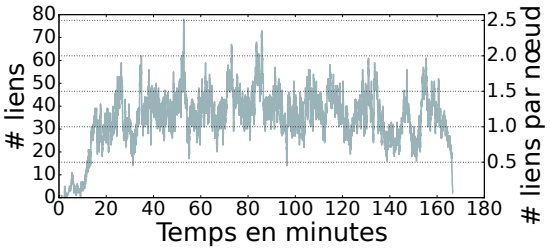
Applications

Jeux de données : plusieurs réseaux d'interactions.

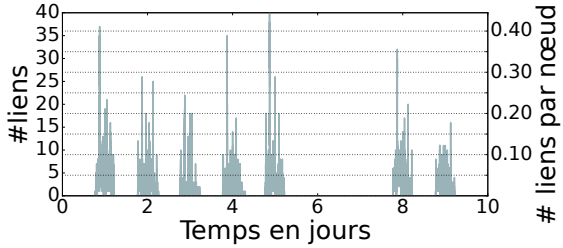
Jeux de données	Nœuds	Liens	Durée
Rollernet	62	15803	3 heures
Socio Pattern	180	19774	9 jours
Reality Mining	94	44975	9 mois
Babouin	28	95616	14 jours

Différences de dynamique

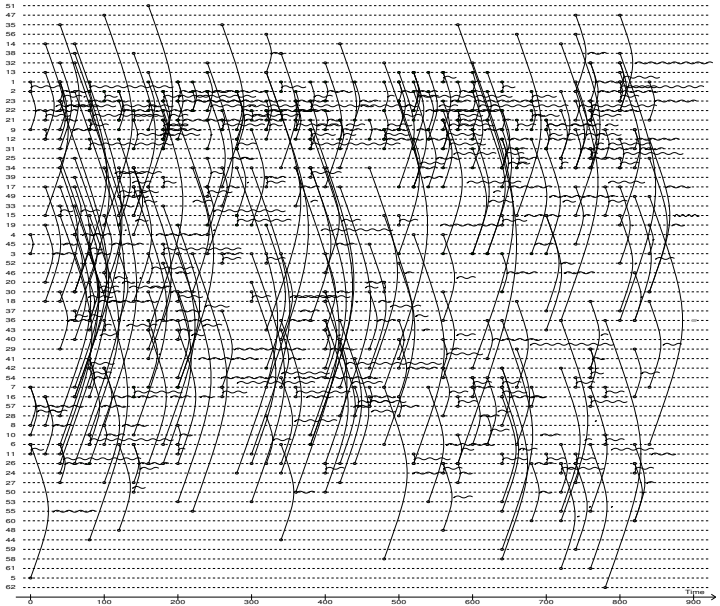
Rollernet



Socio Pattern

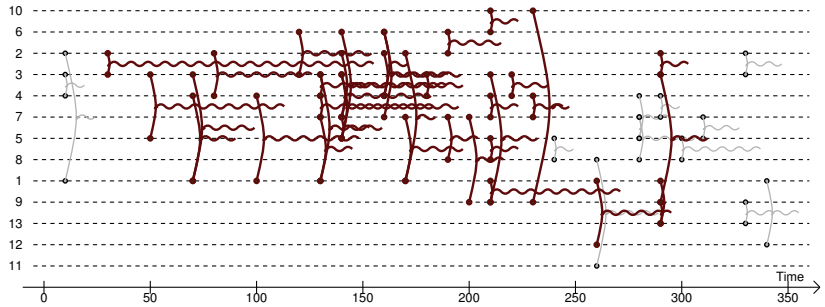


Rollernet



15 minutes du flot de Rollernet.

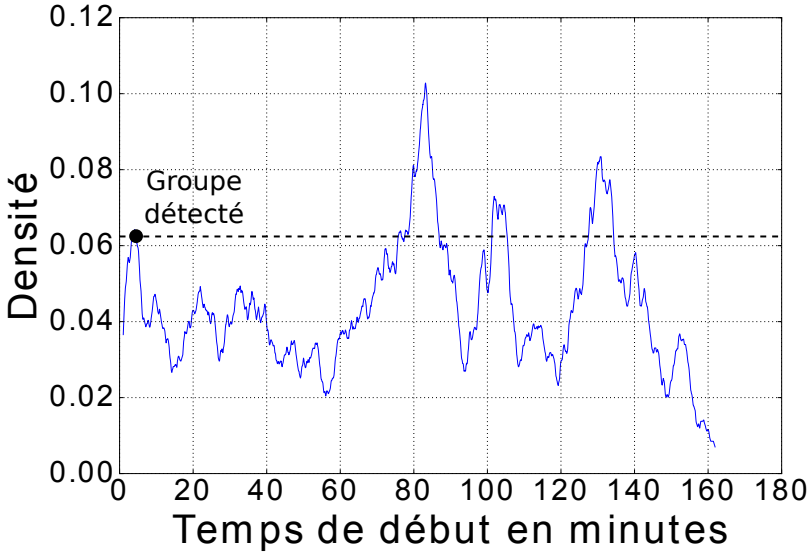
Rollernet : étude d'un groupe



38 interactions entre 11 personnes pendant 5 minutes au début de la randonnée.

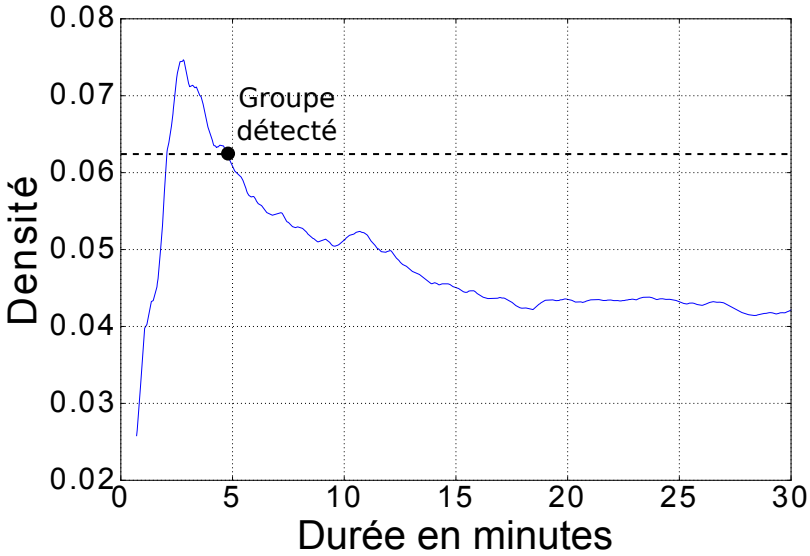
10 personnes étiquetées comme organisateurs à l'arrière de la randonnée

Rollernet : étude d'un groupe



$score = 0.86$

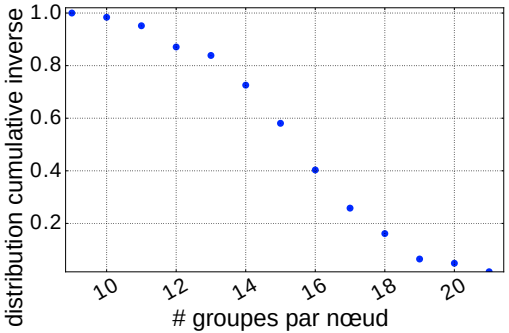
Rollernet : étude d'un groupe



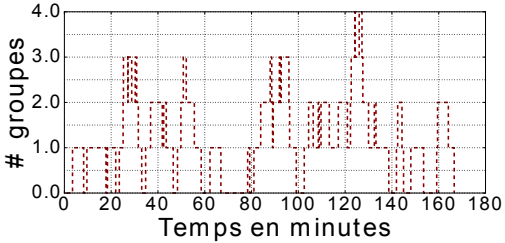
$score = 0.90$

Rollernet : résultats

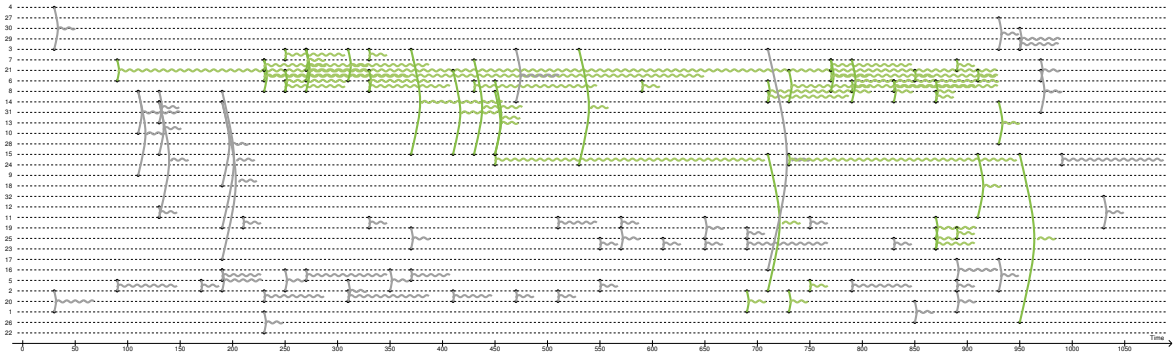
Fort chevauchement topologique



Chevauchement temporel



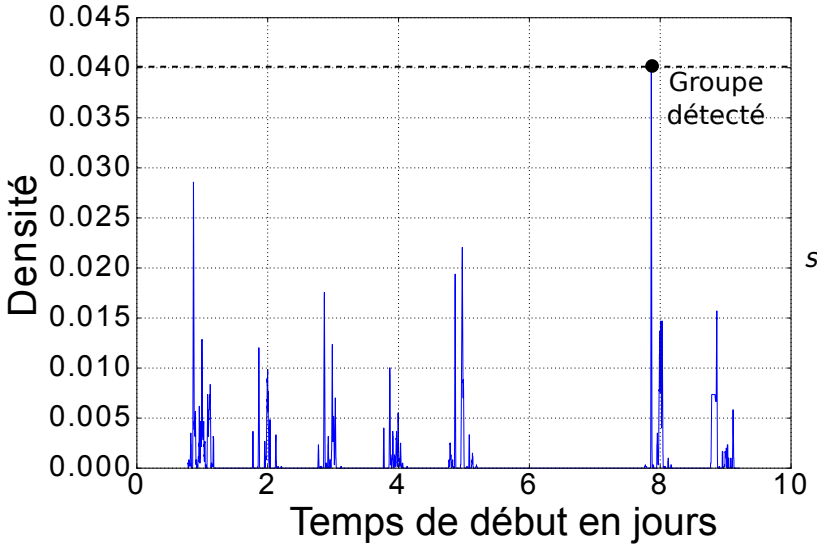
Socio Pattern : étude d'un groupe



50 interactions entre 17 étudiants pendant environ $\approx$ 15 minutes à 7h44.

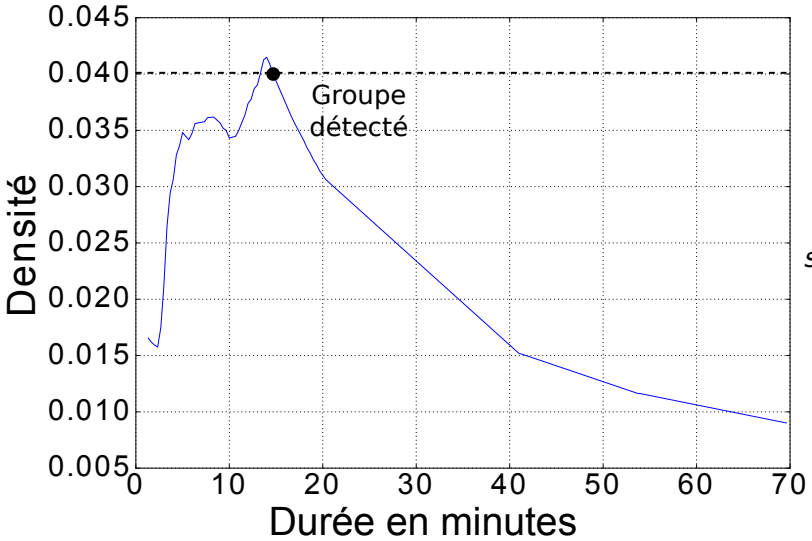
15 personnes de la même classe

Socio Pattern : étude d'un groupe



$score = 1$

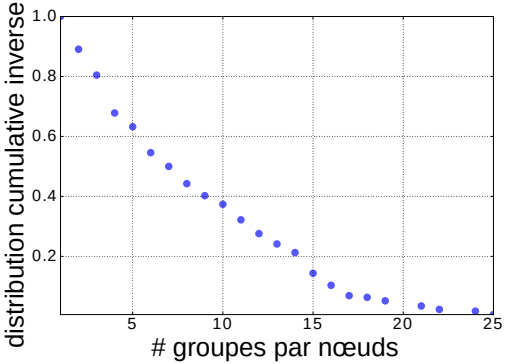
Socio Pattern : étude d'un groupe



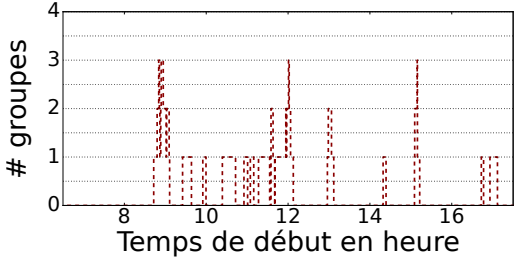
$score = 0.98$

Socio pattern : résultats

Fort chevauchement topologique



Chevauchement temporel



Conclusions

En résumé

- La densité prend en compte le temps et la structure.
- Les références permettent d'évaluer la pertinence d'un groupe de liens.
- De nombreux groupes pertinents détectés dans tous les jeux de données.

Vers des partitions de flots de liens

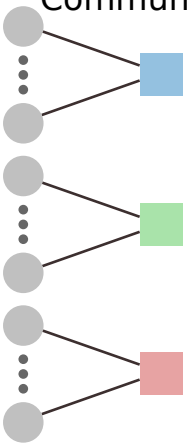
Génération de flots de liens
Modularité temporelle
Tests du générateur
Conclusions

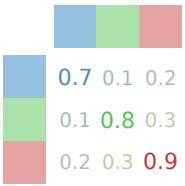
Méthode existante :

Stochastic block model

Nœuds

Communautés

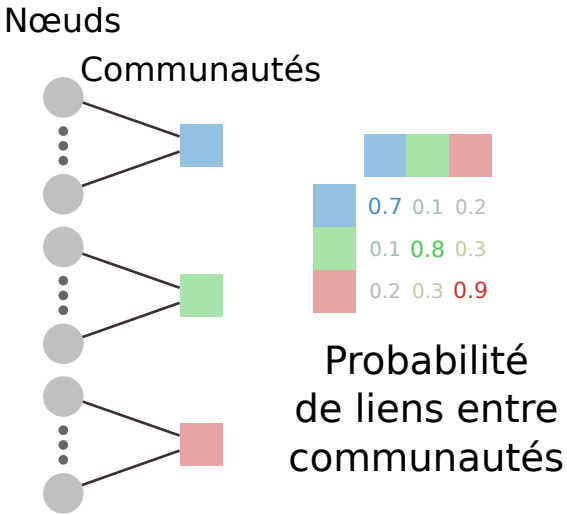


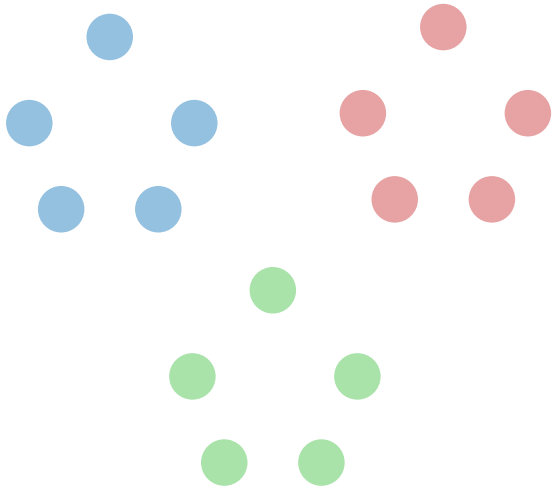


Probabilité de liens entre communautés

Méthode existante :

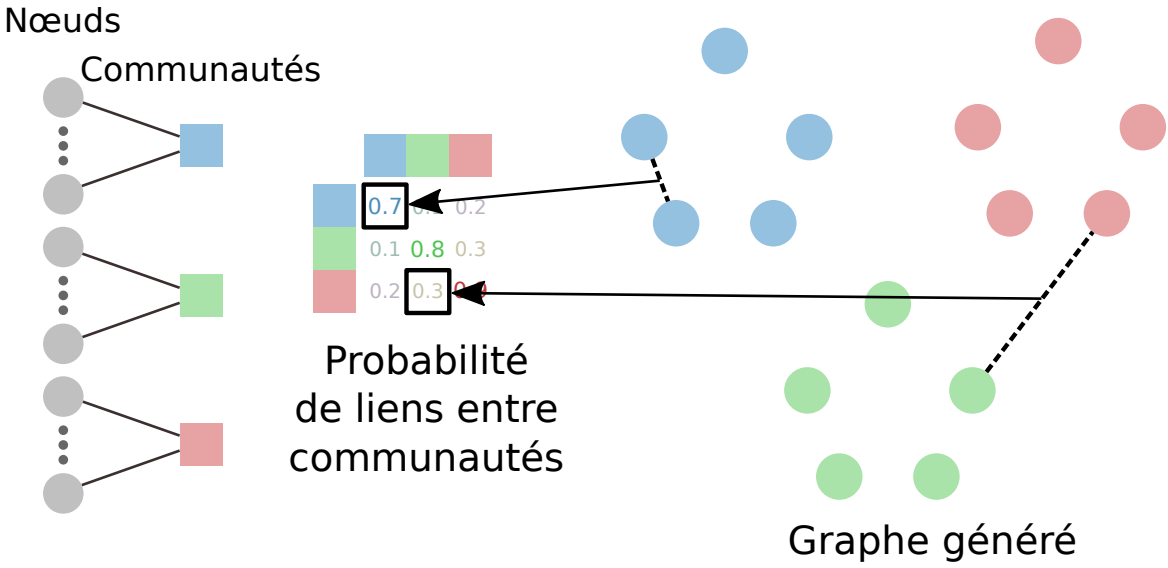
Stochastic block model





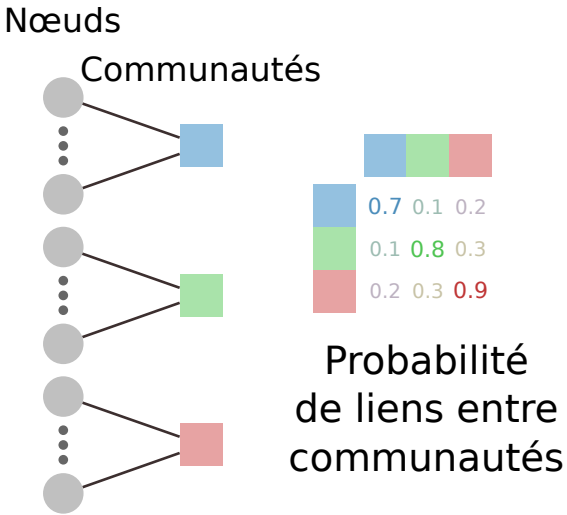
Méthode existante :

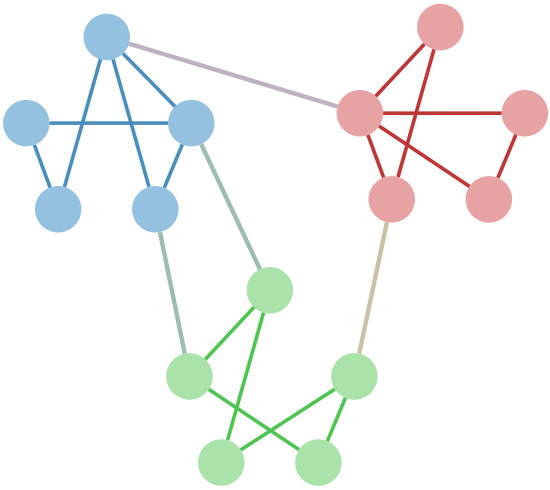
Stochastic block model



Méthode existante :

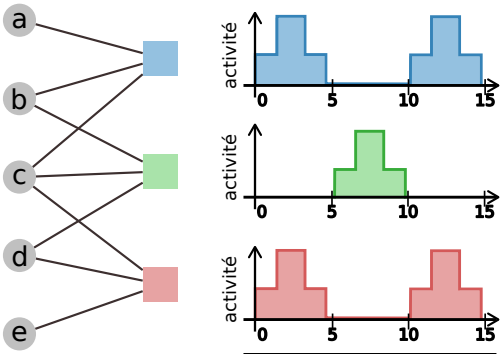
Stochastic block model





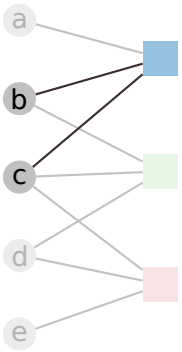
Notre proposition

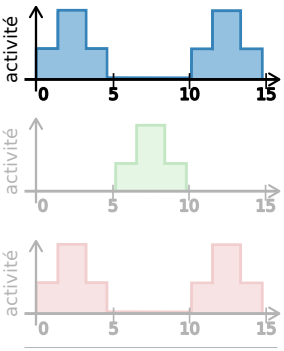
Génération par processus de Poisson non-homogènes indépendants

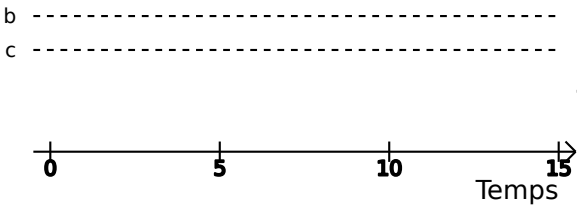


Notre proposition

Génération par processus de Poisson non-homogènes indépendants

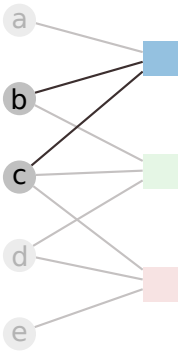


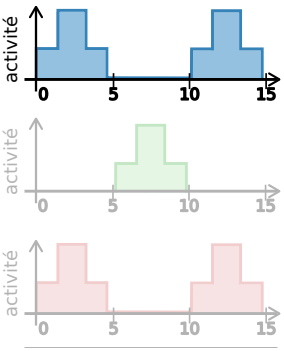


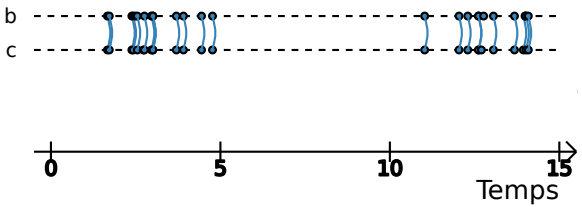


Notre proposition

Génération par processus de Poisson non-homogènes indépendants

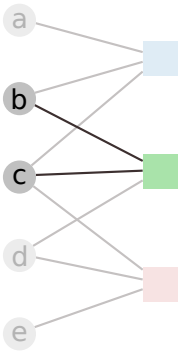


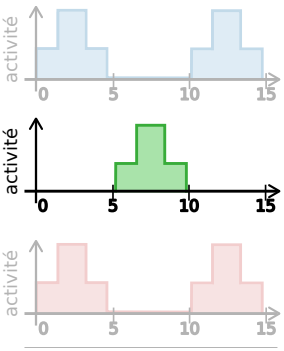


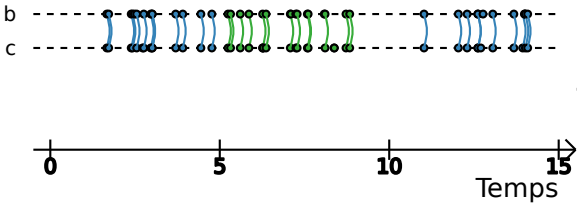


Notre proposition

Génération par processus de Poisson non-homogènes indépendants

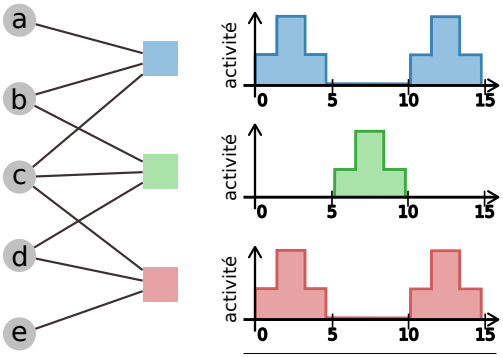


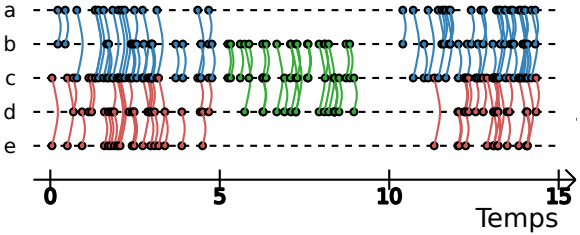




Notre proposition

Génération par processus de Poisson non-homogènes indépendants





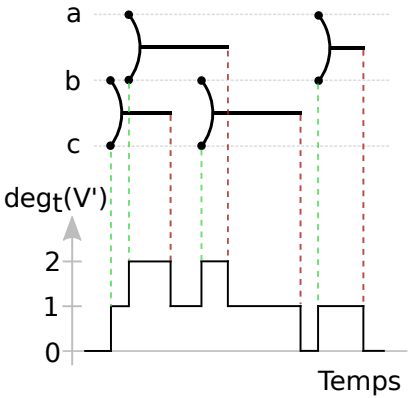
Ajout d'une durée à chaque lien si besoin

Modularité

Modularité d'un ensemble de nœuds V' dans un graphe G :

$$Q_G(V') = f(deg^{in}(V'), deg(V'))$$

Modularité

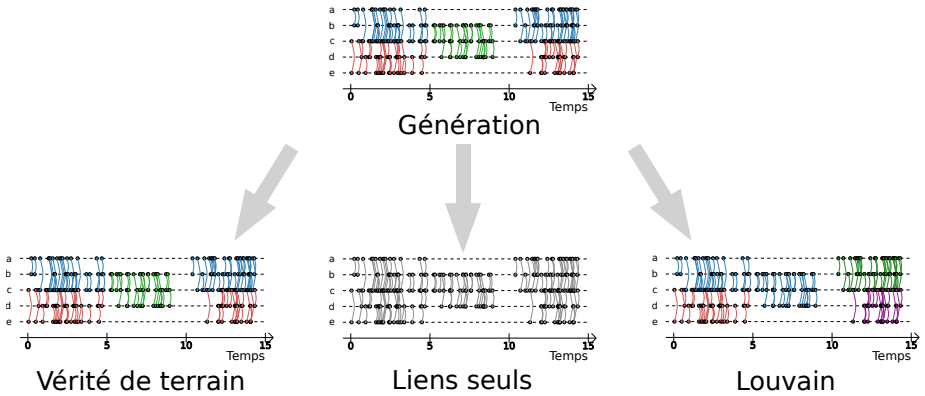


Modularité temporelle de V' dans un flot de liens sur l'intervalle $[\alpha', \omega']$:

$$Q_L(V', \alpha', \omega') = \int_{\alpha'}^{\omega'} f(deg_t^{in}(V'), deg_t(V'))\, dt$$

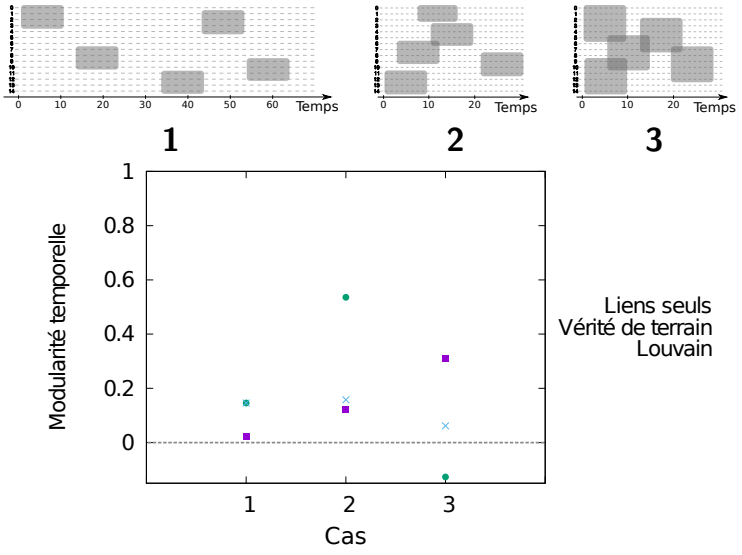
Méthodologie

Génération de flots de liens ayant une structure communautaire de **liens**.



Sur chaque flot de liens, évaluation par la **modularité temporelle** des trois partitions.

Test de trois configurations



Lorsque le chevauchement est trop important, la fonction de qualité n'est plus adaptée.

Conclusions

En résumé

- Génération de nombreuses structures de flots de liens.
- Une première fonction de qualité pour évaluer une partition de liens.
- Méthodes statiques parfois inefficaces pour détecter la vérité de terrain.

Conclusions

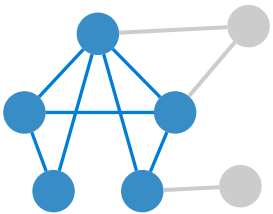
La modélisation sous la forme de flots de liens ...
définition et implémentation

... permet de définir de nouveaux outils de mesure.
description, détection et génération

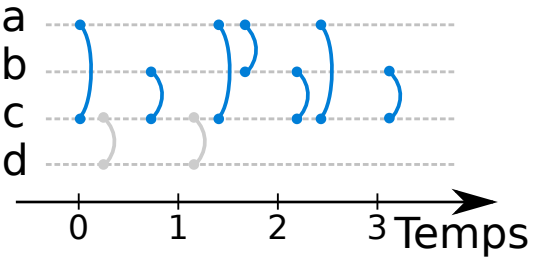
Perspectives

- Extension du formalisme
  - Orientation
  - Poids non constant sur les liens
- Approfondissement des applications
  - Générations de flots de liens réalistes
  - Prédiction de liens
  - Récurrence de groupes pertinents
  - **Fonctions de qualité locales**

Fonctions de qualité locales



cohésion = nombreux chemins internes à la communauté



cohésion =

- nombreux chemins internes à la communauté
- nombreux cycles internes courts
- marcheurs aléatoires "piégés" à l'intérieur de la communauté

Merci !

Expected Nodes

Random graph with the same degree distribution.
Random sampling without replacement of $2|L|$ stub.
B_u : random variable corresponds to how many time u is picked,
$B_u \sim HyperGeom\,(2|E|, d(u), 2|L|)$ [1].

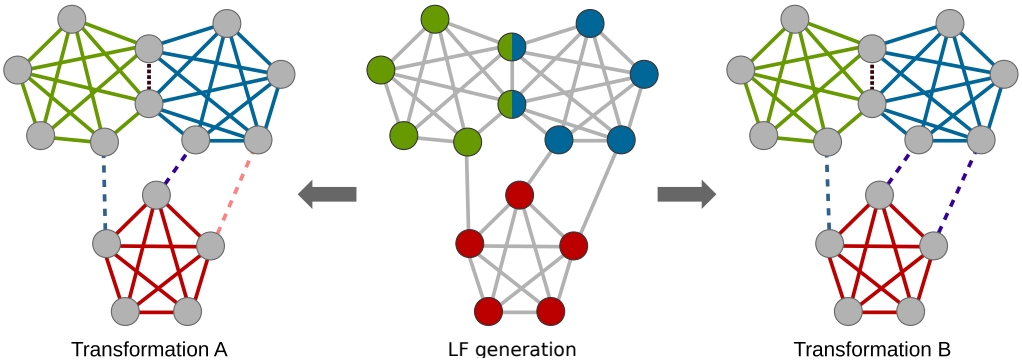
$$\mathbb{P}(B_u = 0) = \frac{\binom{2|E|-d(u)}{2|L|}}{\binom{2|E|}{2|L|}}$$

$$\mathbb{E}[|V(L)|] = \sum_{u \in V} 1 - \mathbb{P}(B_u = 0)\mu_G(m) = \sum_{u \in V} 1 - \frac{\binom{2|E|-d_G(u)}{2m}}{\binom{2|E|}{2m}}$$

1. This can be approximate with a Binomial.

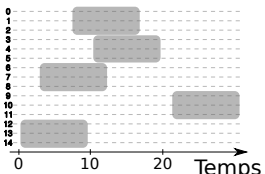
Génération de vérité de terrain

Transformation d'une couverture de nœuds vers une partition de liens

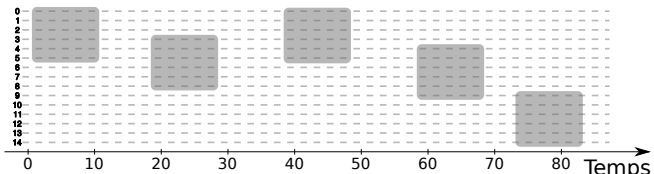


Nombre de nœuds et de communautés	fixe
Durée du flot de liens	**variable**
Durée et intensité de l'activité d'une communauté	fixe
Temps de début d'une communauté	uniforme
Densité du graphe d'affiliation	**variable**

Chevauchements temporel et topologique manipulables séparément



Fort chevauchement temporel



Fort chevauchement topologique

Modularité temporelle

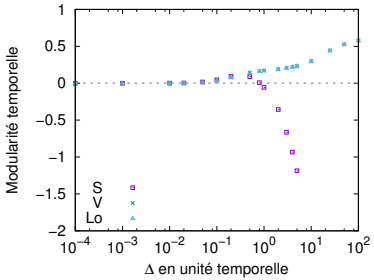
Modularité dans un graphe G d'un ensemble de nœuds V_i

$$Q_G(V_i) = \frac{d_{in}(V_i)}{2m} - \left(\frac{d(V_i)}{2m}\right)^2$$

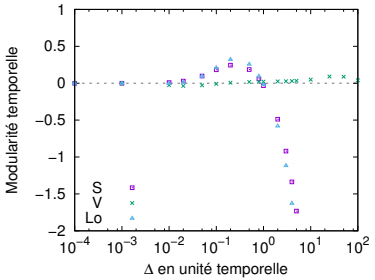
Modularité temporelle dans un flot de liens L d'un ensemble de nœuds V_i sur l'intervalle $[\alpha_i, \omega_i]$

$$Q_L(V_i, \alpha_i, \omega_i) = \int_{\alpha_i}^{\omega_i} \frac{d_{in}(t, V_i)}{d(t, V)} - \left(\frac{d(V_i, t)}{d(V, t)}\right)^2 dt$$

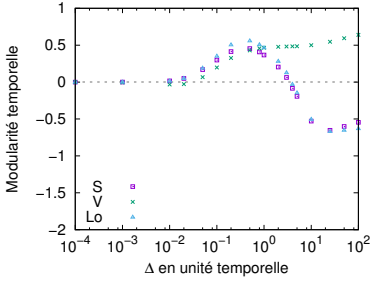
Résultat en fonction de delta



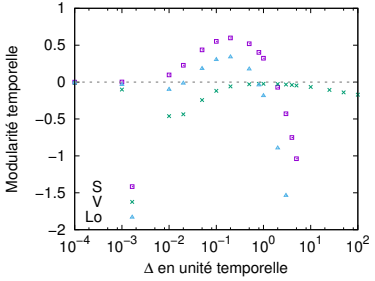
T=200 et O=1.1



T=200 et O=2



T=50 et O=1.1



T=50 et O=2