

MARAMI 2012 17/10/2012

Vincent Labatut

Université Galatasaray – Département d'informatique

[vlabatut@gsu.edu.tr](mailto:vlabatut@gsu.edu.tr) | [bit.gsu.edu.tr/compnet](http://bit.gsu.edu.tr/compnet)

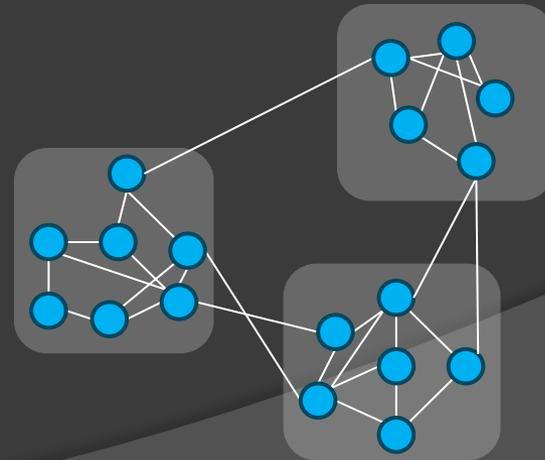
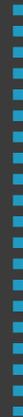
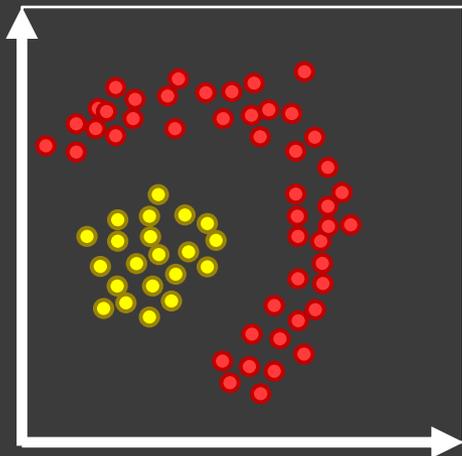
# UNE NOUVELLE MESURE POUR L'ÉVALUATION DES MÉTHODES DE DÉTECTION DE COMMUNAUTÉS

# Plan

- ⦿ Introduction
- ⦿ Mesures couramment utilisées
- ⦿ Limitation de l'approche classique
- ⦿ Nouvelle mesure proposée
- ⦿ Évaluation expérimentale
- ⦿ Conclusion

# Détection de communautés

- ⊙ **Définition** : identification de groupes de nœuds plus densément connectés entre eux qu'avec le reste du réseau
- ⊙ **Structure de communautés** : partition de l'ensemble des nœuds
- ⊙ **Intérêt** : étude du réseau à un niveau intermédiaire
- ⊙ **Problème lié** : partitionnement de données (*cluster analysis*)
  - Même but : former des groupes d'objets
  - Données différentes : information individuelle vs. relationnelle



# Approche typique

- ⦿ Description d'une nouvelle méthode de détection de communautés :
  1. Description de l'algorithme
  2. Application à des réseaux réels et/ou artificiels
  3. Comparaison avec les performances d'autres méthodes
- ⦿ Nombreuses questions méthodologiques :
  - Réseaux réels : sélection, identification des communautés de référence
  - Réseaux artificiels : méthode de génération, paramètres, réalisme
  - Mesure objective de la performance
  - Généralisation à d'autres données

# Comparaison de partitions

- ⊙ Issues du partitionnement de données, ou similaires
- ⊙ Comparaison de deux partitions
  - Estimation :  $\{ \quad \}$
  - Référence :  $\{ \quad \}$
- ⊙ Principales mesures :
  - Proportion de nœuds correctement classifiés = pureté
  - Index de Rand (ajusté)
  - Information mutuelle normalisée

# Mesure de pureté

- Pureté d'une partie par rapport à une partition :

$$\frac{(\quad)}{(\quad)} \quad \frac{\quad}{\quad}$$

- Interprétation : proportion correspondant à la partie de majoritaire dans

- Pureté d'une partition par rapport à une partition :

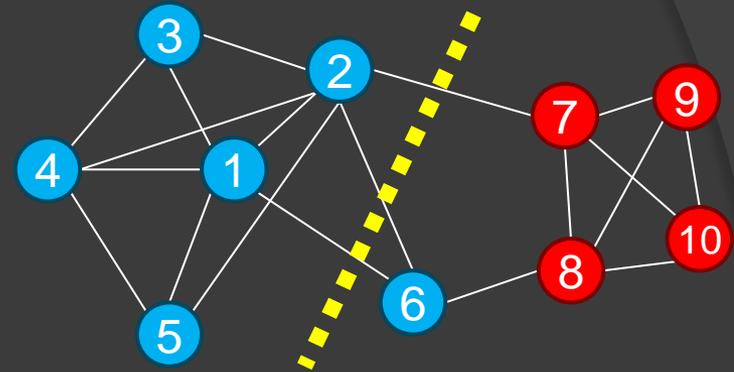
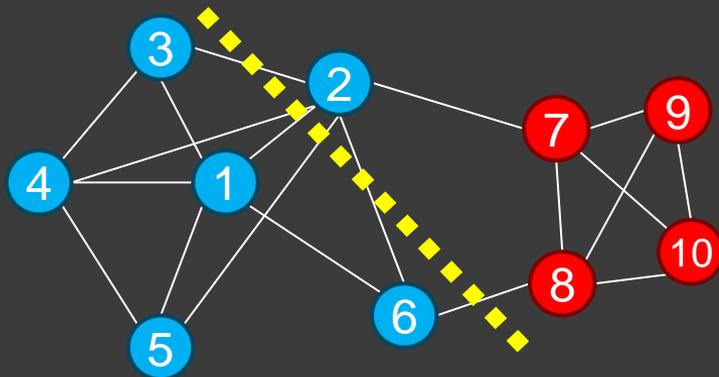
$$\frac{(\quad)}{(\quad)} \quad \frac{\quad}{\quad} \quad (\quad)$$

- : nombre d'éléments dans l'ensemble partitionné
- Interprétation : somme des puretés des classes, pondérée par leurs prévalences
- Sensible au nombre de parties

- F-mesure de deux partitions et :

$$\frac{(\quad)}{(\quad)} \quad \frac{(\quad) \quad (\quad)}{(\quad) \quad (\quad)}$$

# Limitation de ces approches



- ⊙ Exemple :
  - 2 communautés dans la partition de référence
  - 2 estimations et comportant un nœud mal classifié
- ⊙ Pureté :
  - $( ) - - - - ; ( ) - - - - ; ( )$
  - Par définition, mêmes résultats pour
- ⊙ Modularité :
  - $( ) ; ( ) ; ( )$
- ⊙ Remarque valable pour toutes les mesures issues du partitionnement de données

# Pureté nodale

- Pureté d'un nœud  $v$  pour une partition  $\pi$ , par rapport à une partition  $\rho$  :

$$p_{\rho, \pi}(v) = \sum_{\substack{\sigma \in \pi \\ v \in \sigma}} \delta_{\rho, \sigma}$$

- $\delta_{\rho, \sigma}$  ;
- est le symbole de Kronecker :  $\delta_{\rho, \sigma} = 1$  si  $\rho = \sigma$ , 0 sinon
- Interprétation : 1 si la partie de  $\pi$  majoritaire dans la partie de  $\rho$  contenant  $v$  contient elle aussi  $v$  ; 0 sinon
- Ex. :  $p_{\rho, \pi}(v) = \delta_{\rho, \sigma} = 1$  ;  $p_{\rho, \pi}(v) = \delta_{\rho, \tau} = 0$

# Reformulation de la pureté

- Pureté d'une partie , par rapport à une partition : moyenne des puretés nodales

$$\left( \begin{array}{c} \text{ } \\ \text{ } \end{array} \right) \frac{\text{---}}{\text{---}} \left( \begin{array}{c} \text{ } \\ \text{ } \end{array} \right)$$

- Pureté d'une partition par rapport à une partition : on applique la formule initiale

$$\left( \begin{array}{c} \text{ } \\ \text{ } \end{array} \right) \frac{\text{---}}{\text{---}} \left( \begin{array}{c} \text{ } \\ \text{ } \end{array} \right) - \left( \begin{array}{c} \text{ } \\ \text{ } \end{array} \right)$$

- Expression équivalente à la première définition de la pureté
- Intérêt : pureté exprimée de façon individuelle

# Mesure proposée

- Pureté modifiée :

$$\left( \quad \right) \quad \text{—} \quad \left( \quad \right)$$

- Poids représentant l'importance topologique du nœud

- Proposition de poids :

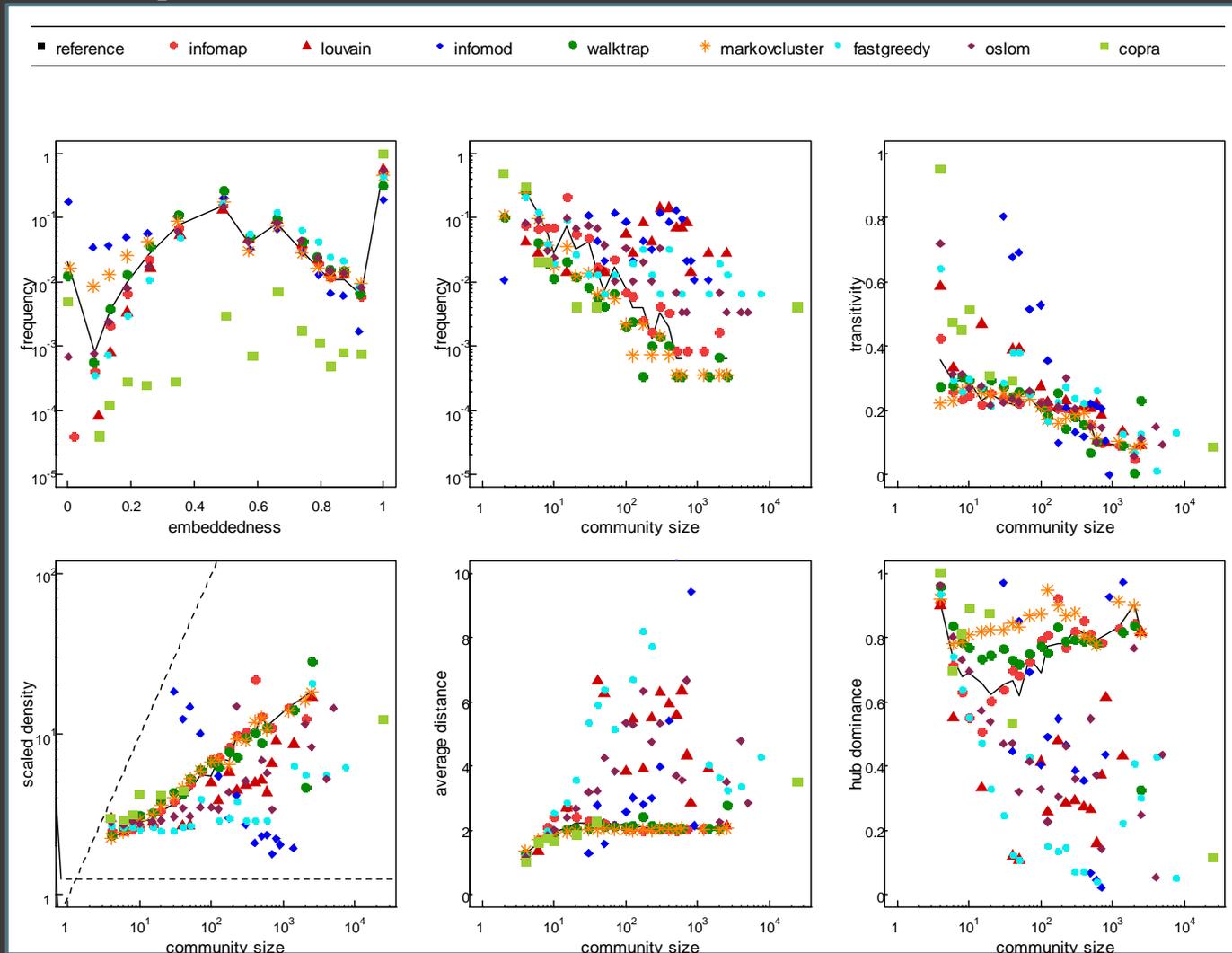
- Degré normalisé :
- Enchâssement :
- Combinaison des deux :

- F-mesure modifiée :

$$\left( \quad \right) \quad \frac{\left( \quad \right) \quad \left( \quad \right)}{\left( \quad \right) \quad \left( \quad \right)}$$

- Ex. :  $\left( \quad \right)$  ;  $\left( \quad \right)$

# Étude préliminaire

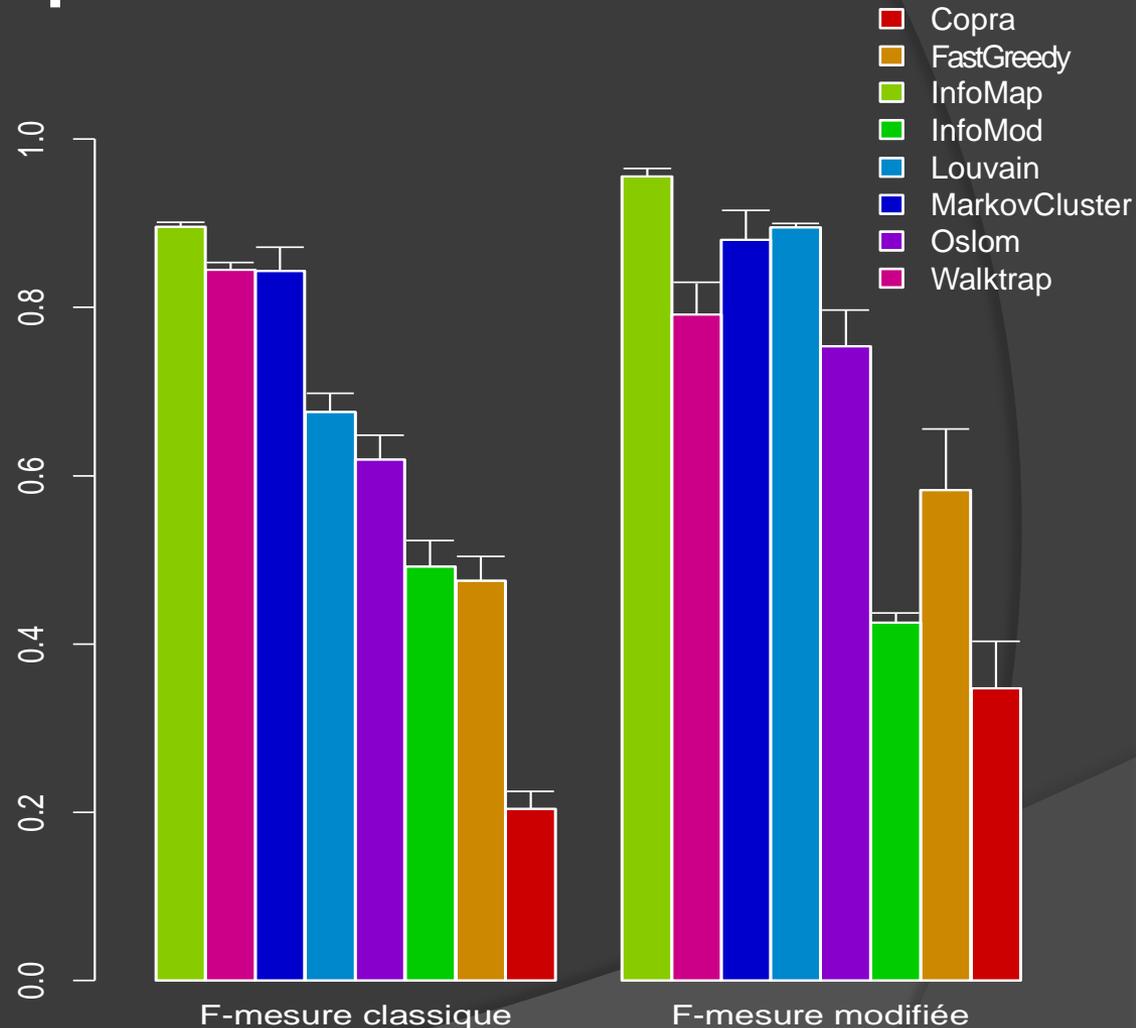


[Orman et al. 2012]

# Évaluation expérimentale

F-mesure classique	
Rang	Algorithmes
1	InfoMap
2-3	MarkovCluster, WalkTrap
4-5	Louvain, Osлом
6-7	InfoMod, FastGreedy
8	Copra

F-mesure modifiée	
Rang	Algorithmes
1	InfoMap
2-3	<b>Louvain</b> , MarkovCluster
4-5	<b>WalkTrap</b> , Osлом
6	FastGreedy
7-8	<b>InfoMod</b> , Copra

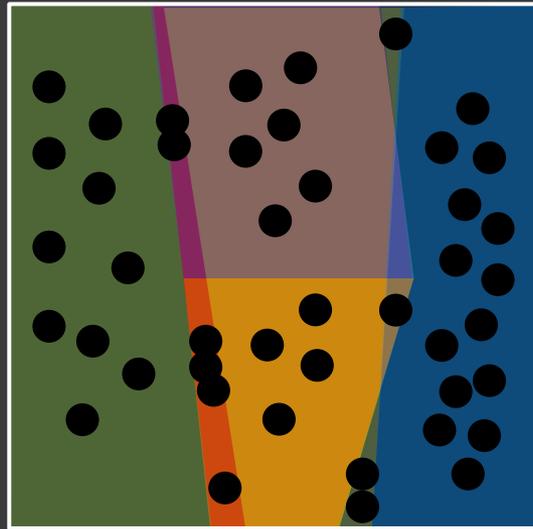
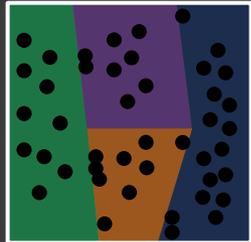
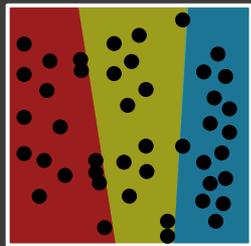


# Conclusion

- ⦿ Contributions
  - Mise en évidence de la limitation des mesures de performance
  - Proposition d'une solution
  - Test sur des données contrôlées
- ⦿ Limitations
  - Effet limité sur les classements
  - Définition du poids
  - Correction de la mesure
- ⦿ Extensions
  - Comparaison de partitions estimées
  - Application à la partition de données
  - Adaptation des poids aux particularités du réseau
  - Application du même principe à IMN et index de Rand

# Annexes

# Comparaison de deux partitions



	Green	Purple	Orange	Dark Blue
Red	10	2	4	0
Yellow	0	6	4	3
Light Blue	0	0	1	14