Identifier les rôles communautaires des capitalistes sociaux sur Twitter

Nicolas Dugué[1] Vincent Labatut[2] Anthony Perez[1]

[1]Université d'Orléans, LIFO
{nicolas.dugue,anthony.perez}@univ-orleans.fr

[2]Université Galatasaray, Département d'informatique
vlabatut@gsu.edu.tr

AMLaGAP
19 mai 2014

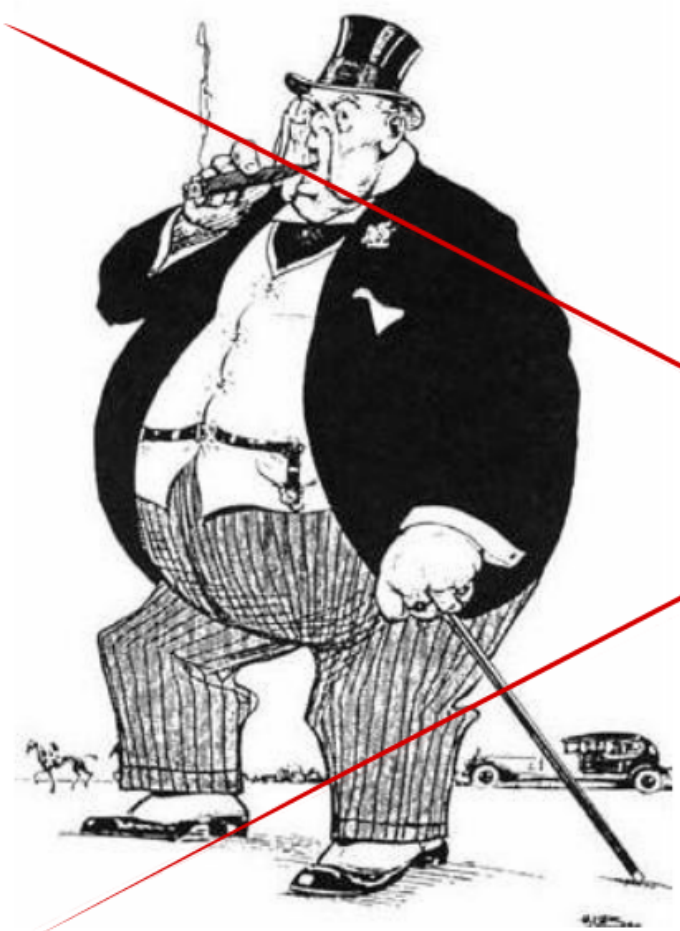
Plan

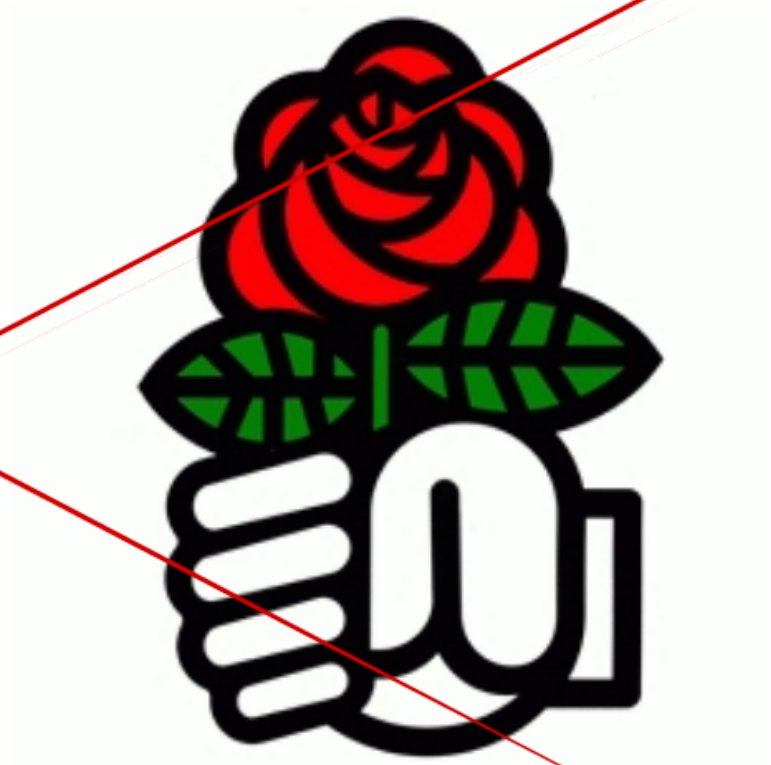
twitter


Jure Leskovec
@jure
Professor of #computerscience @Stanford. Thinking about #datamining massive social and information #networks, #bigdata, #web and #socialmedia.
Stanford, CA · cs.stanford.edu/~jure/
TWEETS
421
ABONNEMENTS
121
ABONNÉS
7 191
Abonné

Capitalistes sociaux ?

Capitalistes sociaux ?





Capitalistes sociaux ?

Capital social

Bourdieu : ”l’ensemble des ressources actuelles ou potentielles qui sont liées à la possession d’un réseau durable de relations”

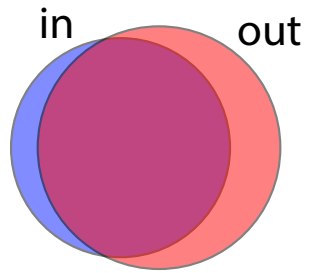
Notion de capitalisme social [GVK+12]

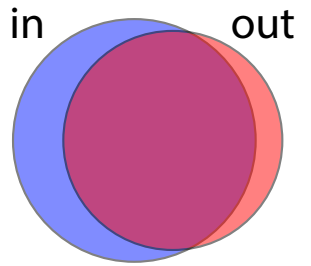
Introduite par Ghosh et al.[GVK+12] : Utilisateurs qui suivent le plus les spammeurs.
Liste de 100.000 utilisateurs.

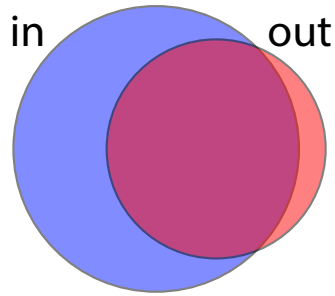
- Obtenir rapidement un maximum de visibilité
- Pourquoi les étudier ?
  - Comprendre leur influence sur le réseau
  - Améliorer la qualité de service
  - Appliquer leurs méthodes à d'autres domaines (marketing)

Stratégies des capitalistes sociaux

- I Follow You, Follow Me (IFYFM)
- Follow Me, I Follow You (FMIFY)
- État passif







Exemples de capitalistes sociaux

screen name	name	followers	friends
IFOLLOWBACKJP	TFBJP	$1.2 \cdot 10^5$	$1.1 \cdot 10^5$
itsrealchris	iFollowBack	$1.7 \cdot 10^5$	$1.6 \cdot 10^5$
AllFollowMax	TFBJP	$4.2 \cdot 10^4$	$4.3 \cdot 10^4$
BarackObama	Barack Obama	$2.5 \cdot 10^7$	$6.7 \cdot 10^5$
britneyspears	Britney Spears	$2.2 \cdot 10^7$	$4.1 \cdot 10^5$
JetBlue	JetBlue Airways	$1.7 \cdot 10^6$	$1.1 \cdot 10^5$
Starbucks	Starbucks Coffee	$3.2 \cdot 10^6$	$7.9 \cdot 10^4$

Table: Followers and friends numbers are rounded.

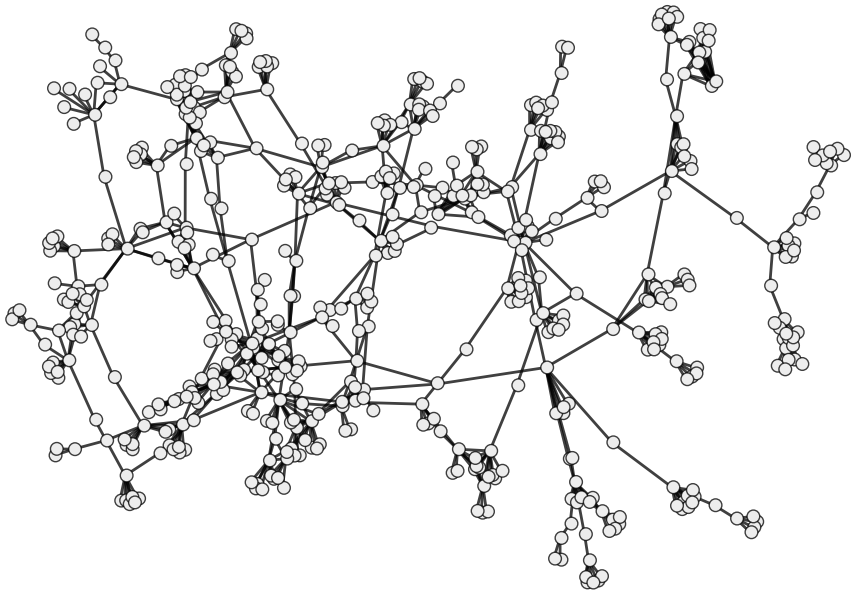
Plan

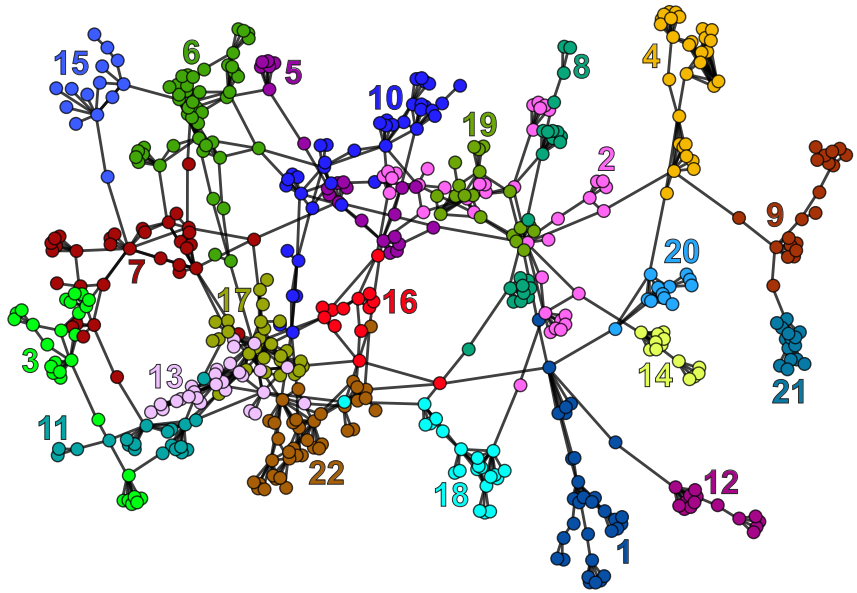
Structure de communautés

Définition

Partition du graphe telle que les noeuds d'une partie sont plus connectés entre eux qu'avec le reste du graphe.

Niveau d'observation intermédiaire.





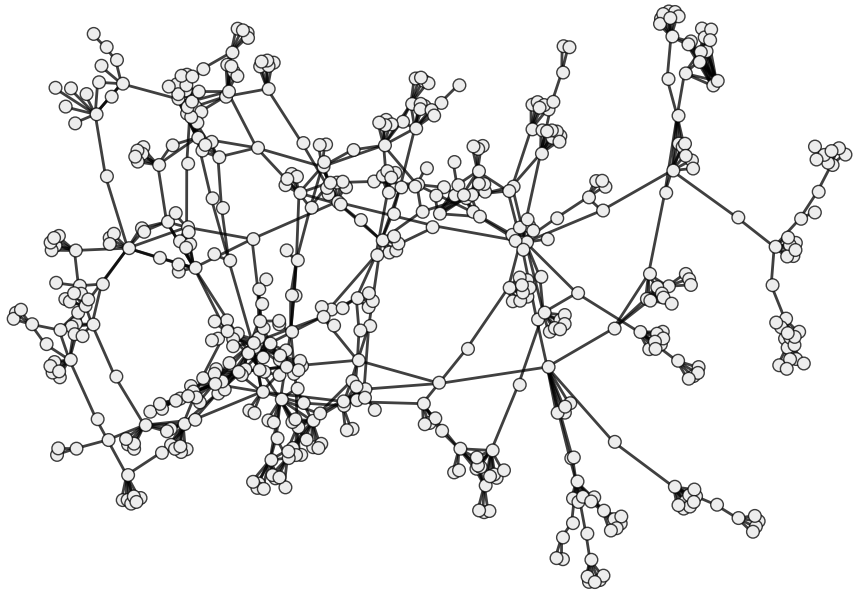
Rôle communautaire

- Sont ils liés avec des communautés extérieures ?
- Sont ils hubs dans les communautés ?
- Sont ils isolés dans certaines communautés ?
- Qui suivent ils ?

$\longrightarrow$ Sont ils visibles ?

Méthode de Guimerà & Amaral [GA05]

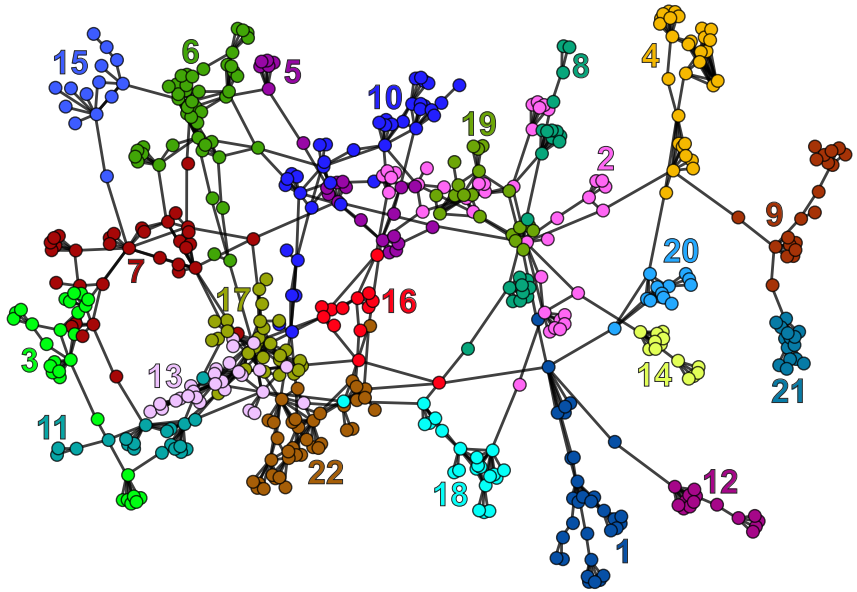
- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés
  2. Calcul des 2 mesures nodales
  3. Partition de l'espace $2D$ obtenu
  4. Mise en correspondance des rôles



[LB12]

Méthode de Guimerà & Amaral [GA05]

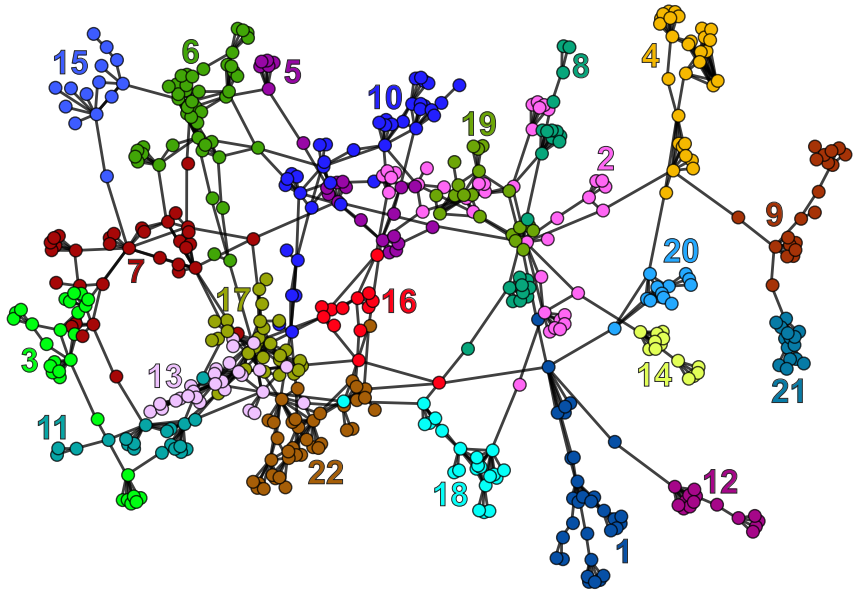
- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés
  2. Calcul des 2 mesures nodales
  3. Partition de l'espace $2D$ obtenu
  4. Mise en correspondance des rôles



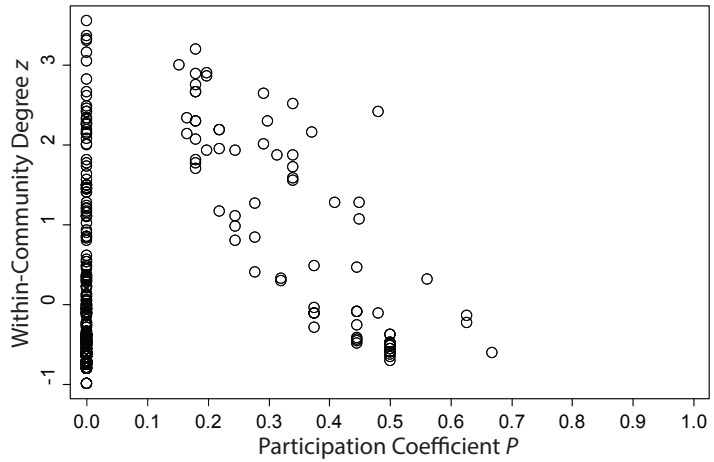
[LB12]

Méthode de Guimerà & Amaral [GA05]

- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés
  2. Calcul des 2 mesures nodales
  3. Partition de l'espace $2D$ obtenu
  4. Mise en correspondance des rôles

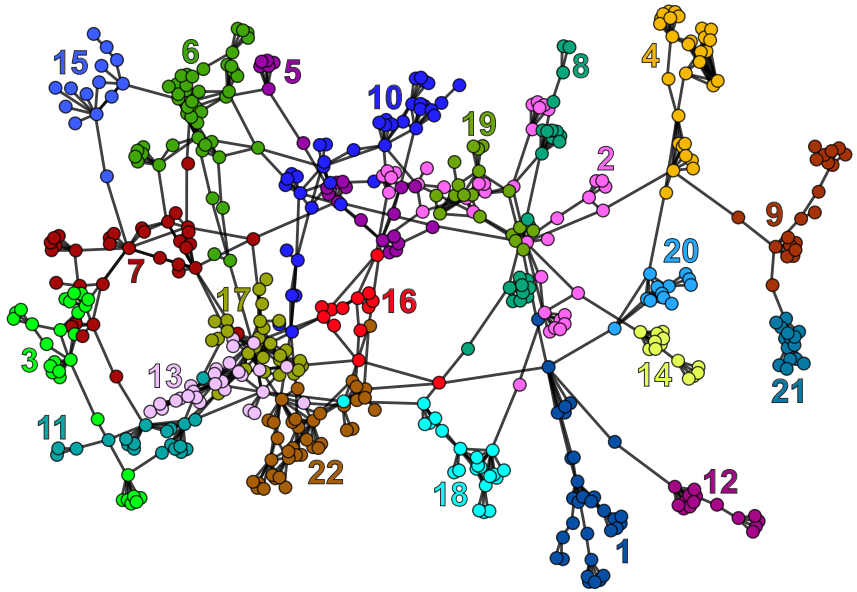


[LB12]

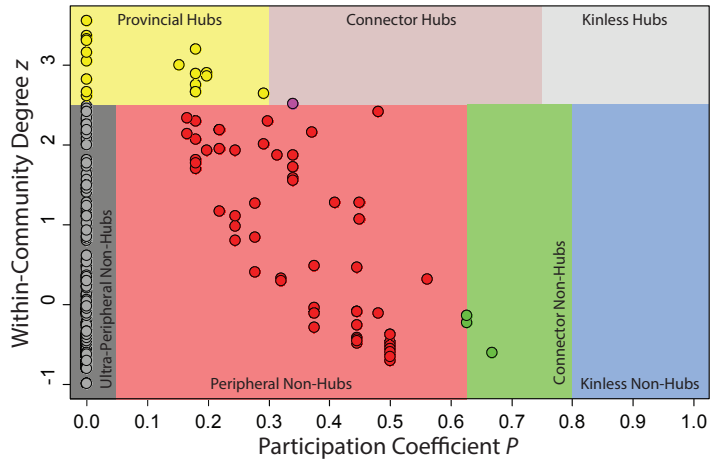


Méthode de Guimerà & Amaral [GA05]

- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés
  2. Calcul des 2 mesures nodales
  3. Partition de l'espace $2D$ obtenu
  4. Mise en correspondance des rôles

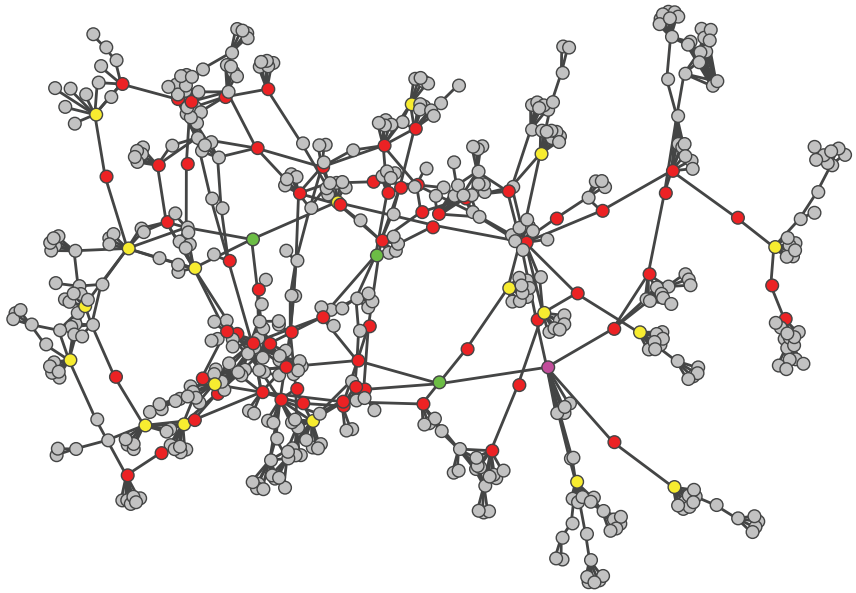


[LB12]

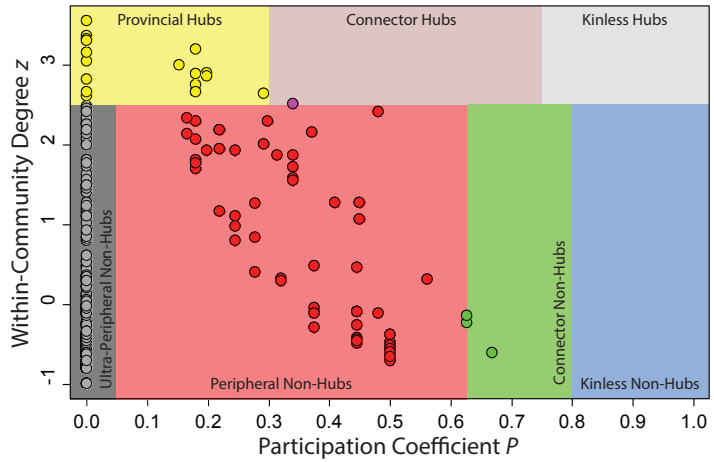


Méthode de Guimerà & Amaral [GA05]

- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés
  2. Calcul des 2 mesures nodales
  3. Partition de l'espace $2D$ obtenu
  4. Mise en correspondance des rôles

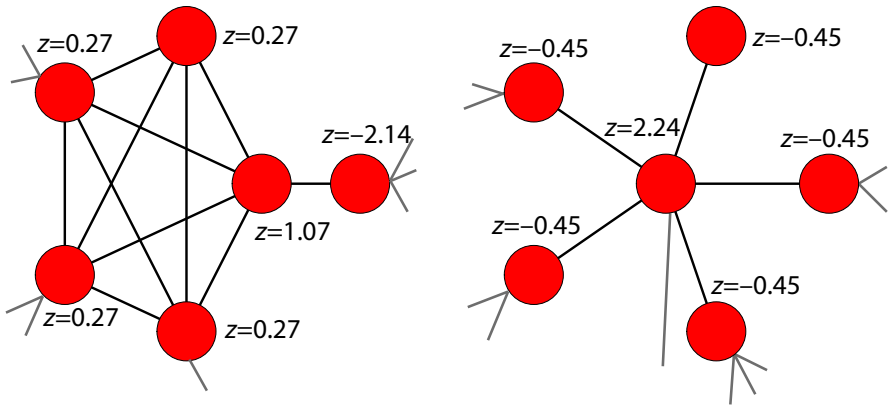


[LB12]



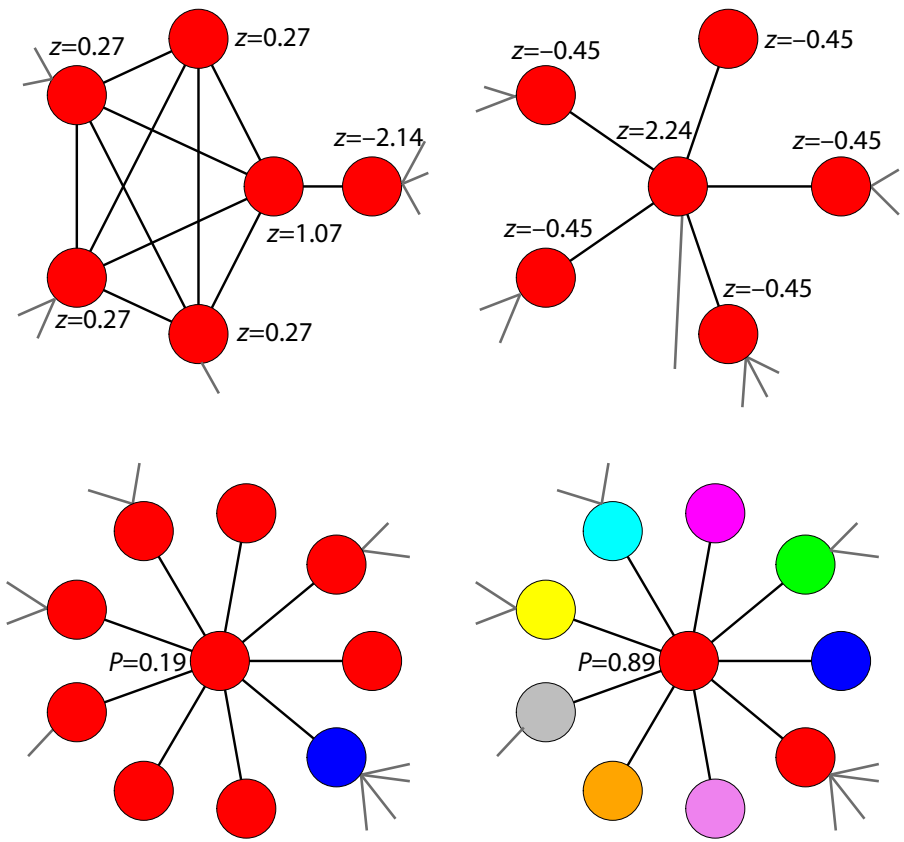
Mesures de rôle

- **Degré interne normalisé**
  - Connectivité *interne*
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score du degré interne k_{int}
  - Bornes pas fixées
- **Coefficient de participation**
  - Connectivité *externe*
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)}\right)^2$
  - k_i : degré pour C_i
  - $P(u) = 0$:
    - Une seule communauté
  - $P(u) \approx 1$:
    - Nombreuses communautés
    - Même nombre de liens



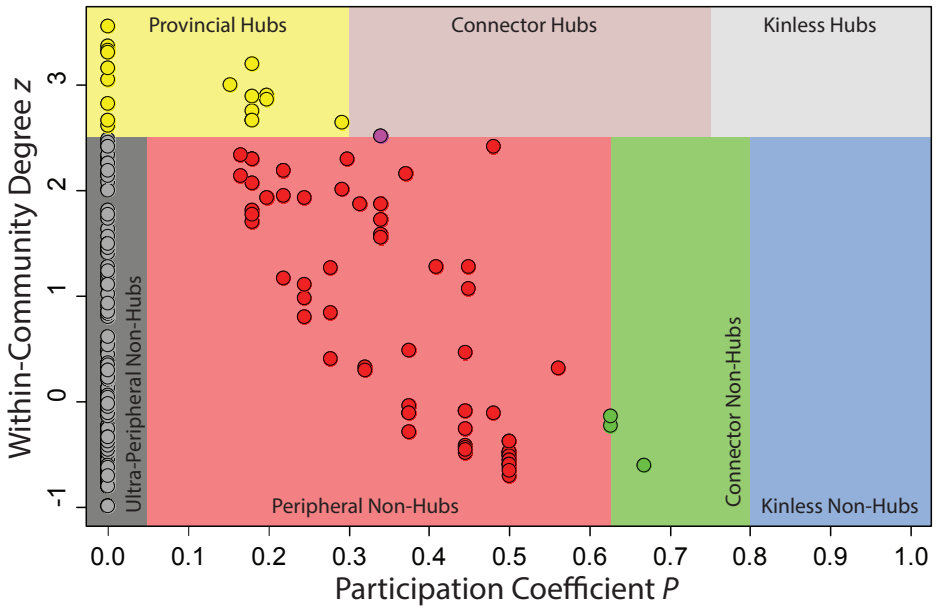
Mesures de rôle

- **Degré interne normalisé**
  - Connectivité *interne*
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score du degré interne k_{int}
  - Bornes pas fixées
- **Coefficient de participation**
  - Connectivité *externe*
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)}\right)^2$
  - k_i : degré pour C_i
  - $P(u) = 0$:
    - Une seule communauté
  - $P(u) \approx 1$:
    - Nombreuses communautés
    - Même nombre de liens



Mesures de rôle

- **Degré interne normalisé**
  - Connectivité *interne*
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score du degré interne k_{int}
  - Bornes pas fixées
- **Coefficient de participation**
  - Connectivité *externe*
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)}\right)^2$
  - k_i : degré pour C_i
  - $P(u) = 0$:
    - Une seule communauté
  - $P(u) \approx 1$:
    - Nombreuses communautés
    - Même nombre de liens

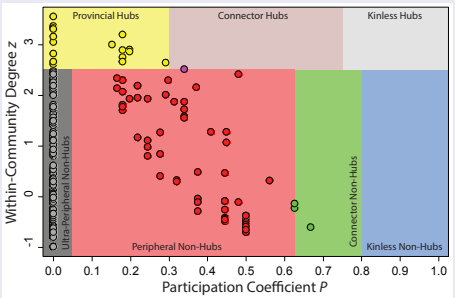


Plan

Limitations de l'approche

Orientation des liens ignorée

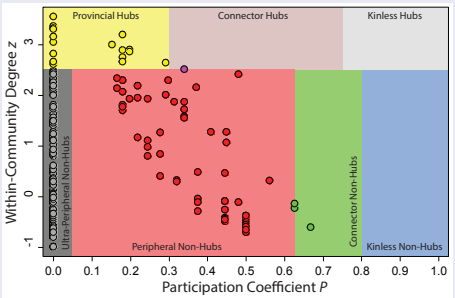
- Systèmes à relations asymétriques
- Twitter : followers / followees
  $\rightarrow$ Quels seuils utiliser en orienté ?



Limitations de l'approche

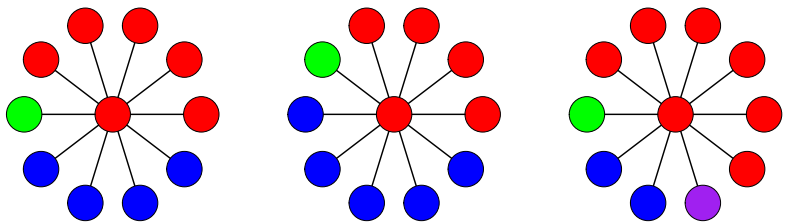
Orientation des liens ignorée

- Systèmes à relations asymétriques
- Twitter : followers / followees
  $\rightarrow$ Quels seuils utiliser en orienté ?



Imprécision de la mesure de connectivité externe

- Degré, nombre de communautés, distribution des liens
- Liens externes, mais aussi internes



$P = 0.58$

Connectivité externe : nouvelle approche

- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité** H
    - Dispersion des liens externes
    - $\lambda(u)$: écart type de k_i
    - $H(u)$: z-score de λ

Connectivité externe : nouvelle approche

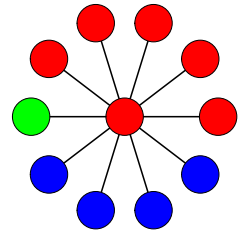
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité** H
    - Dispersion des liens externes
    - $\lambda(u)$: écart type de k_i
    - $H(u)$: z-score de λ

Connectivité externe : nouvelle approche

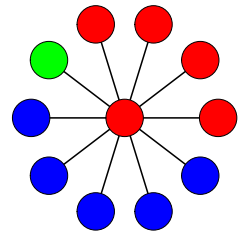
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité** H
    - Dispersion des liens externes
    - $\lambda(u)$: écart type de k_i
    - $H(u)$: z-score de λ

Connectivité externe : nouvelle approche

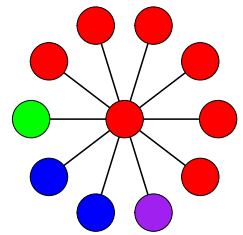
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité** H
    - Dispersion des liens externes
    - $\lambda(u)$: écart type de k_i
    - $H(u)$: z-score de λ



$\epsilon = 2, k_{ext} = 5, \lambda = 1.5$



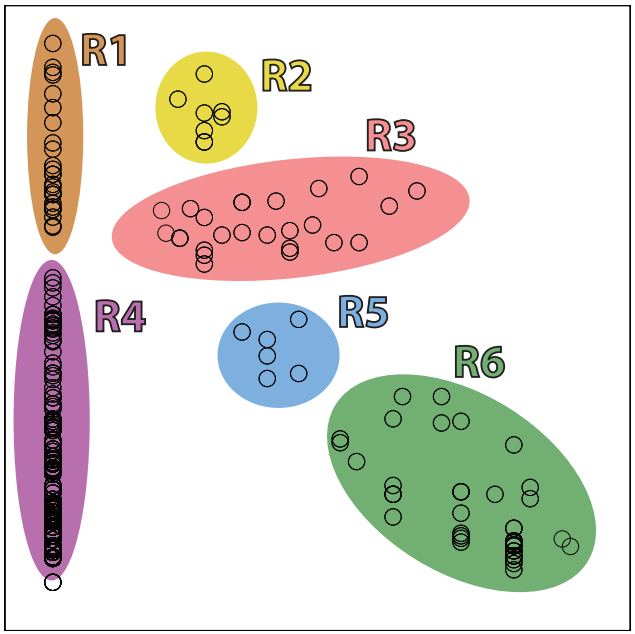
$\epsilon = 2, k_{ext} = 6, \lambda = 2$



$\epsilon = 3, k_{ext} = 4, \lambda = 0.5$

Identification non-supervisée des rôles

- Appliquée à toutes les mesures simultanément
- Chaque groupe obtenu correspond à un rôle



Plan

Réseau étudié

- Collecté en 2009 [CHBG10]
- 55 millions de nœuds (utilisateurs)
- 2 milliards de liens orientés (follower $\rightarrow$ friend)

Méthodologie

- Détection de communautés : Louvain
- Calcul des 8 mesures
- Analyse de regroupement : k-moyennes distribué [Lia09]
- Sélection des groupes : indice Davies-Bouldin $\rightarrow$ méthode du coude

Propriétés des groupes

Groupe	Taille	Proportion	Rôle
1	24543667	$46,68\%$	Non-Hub ultra-périphérique
2	304	$< 0,01\%$	Hub très connecteur (entrant)
3	303674	$0,58\%$	Hub connecteur
4	11929722	$22,69\%$	Non-Hub périphérique (entrant)
5	10828599	$20,59\%$	Non-Hub périphérique (sortant)
6	4973717	$9,46\%$	Non-Hub connecteur

Taille des groupes

Répartition dans les groupes

Ratio	G1 NU	G2 HTC	G3 HC	G4 NP(E)	G5 NP(S)	G6 NC
< 1	$0,03\%$	$0,00\%$	$\mathbf{14,64\%}$	$11,53\%$	$\mathbf{13,65\%}$	$\mathbf{60,15\%}$
	$< 0,01\%$	$0,00\%$	$\mathbf{4,29\%}$	$0,09\%$	$0,11\%$	$1,07\%$
> 1	$0,03\%$	$0,00\%$	$\mathbf{19,38\%}$	$0,48\%$	$\mathbf{14,07\%}$	$\mathbf{66,05\%}$
	$< 0,01\%$	$0,00\%$	$\mathbf{7,31\%}$	$< 0,01\%$	$0,14\%$	$1,52\%$

Capitalistes de faible degré

Ratio	G1	G2	G3	G4	G5	G6
$< 0,7$	$0,00\%$	$10,43\%$	$\mathbf{81,67\%}$	$0,00\%$	$0,00\%$	$7,90\%$
	$0,00\%$	$\mathbf{31,25\%}$	$0,24\%$	$0,00\%$	$0,00\%$	$< 0,01\%$
$> 0,7$ et < 1	$0,00\%$	$1,52\%$	$\mathbf{95,72\%}$	$0,00\%$	$0,00\%$	$2,76\%$
	$0,00\%$	$7,24\%$	$0,46\%$	$0,00\%$	$0,00\%$	$< 0,01\%$
> 1	$0,00\%$	$0,03\%$	$\mathbf{98,02\%}$	$0,00\%$	$0,00\%$	$1,96\%$
	$0,00\%$	$0,33\%$	$1,24\%$	$0,00\%$	$0,00\%$	$< 0,01\%$

Capitalistes de degré élevé

Répartition dans les groupes

Ratio	**G1** NU	**G2** HTC	**G3** HC	**G4** NP(E)	**G5** NP(S)	**G6** NC
< 1	$0,03\%$	$0,00\%$	$\mathbf{14,64\%}$	$11,53\%$	$\mathbf{13,65\%}$	$\mathbf{60,15\%}$
	$< 0,01\%$	$0,00\%$	$\mathbf{4,29\%}$	$0,09\%$	$0,11\%$	$1,07\%$
> 1	$0,03\%$	$0,00\%$	$\mathbf{19,38\%}$	$0,48\%$	$\mathbf{14,07\%}$	$\mathbf{66,05\%}$
	$< 0,01\%$	$0,00\%$	$\mathbf{7,31\%}$	$< 0,01\%$	$0,14\%$	$1,52\%$

Capitalistes de faible degré

Ratio	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
$< 0,7$	$0,00\%$	$10,43\%$	$\mathbf{81,67\%}$	$0,00\%$	$0,00\%$	$7,90\%$
	$0,00\%$	$\mathbf{31,25\%}$	$0,24\%$	$0,00\%$	$0,00\%$	$< 0,01\%$
$> 0,7$ et < 1	$0,00\%$	$1,52\%$	$\mathbf{95,72\%}$	$0,00\%$	$0,00\%$	$2,76\%$
	$0,00\%$	$7,24\%$	$0,46\%$	$0,00\%$	$0,00\%$	$< 0,01\%$
> 1	$0,00\%$	$0,03\%$	$\mathbf{98,02\%}$	$0,00\%$	$0,00\%$	$1,96\%$
	$0,00\%$	$0,33\%$	$1,24\%$	$0,00\%$	$0,00\%$	$< 0,01\%$

Capitalistes de degré élevé

Observation sur le positionnement des capitalistes sociaux

Présence dans des groupes bien spécifiques

- Hubs : G2 et G3
- Connecteurs : G3 et G6
- Très connecteurs : G2

Conclusion

- Contributions
  - Meilleure caractérisation de la connectivité externe
  - Détection automatique des rôles
  - Visibilité des capitalistes sociaux
- Perspectives
  - Visibilité $\rightarrow$ Influence ?
  - Stabilité de la méthode ?
  - Universalité des rôles ?

Références

[CHBG10] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi.

Measuring user influence in twitter: The million follower fallacy.

In *international AAAI Conference on Weblogs and Social Media*, 2010.

[GA05] R. Guimerà and L. Amaral.

Functional cartography of complex metabolic networks.

Nature, 433:895–900, 2005.

[GVK+12] Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi.

Understanding and combating link farming in the twitter social network.

In *21st International Conference on WWW*, pages 61–70, 2012.

Références

[LB12] Vincent Labatut and Jean-Michel Balasque.

Detection and interpretation of communities in complex networks: Methods and practical application.

In Ajith Abraham and Aboul-Ella Hassanien, editors, *Computational Social Networks: Tools, Perspectives and Applications*, chapter 4, pages 81–113. Springer, 2012.

[Lia09] Wei-Keng Liao.

Parallel k-means data clustering, Oct 2009.