Rôle communautaire des capitalistes sociaux dans 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

MARAMI
Jeudi 17 octobre 2013

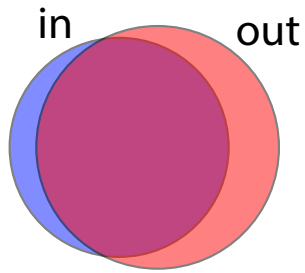
Plan

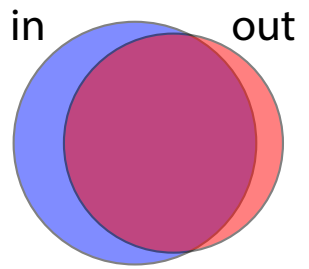
Notion de capitalisme social [GVK+12]

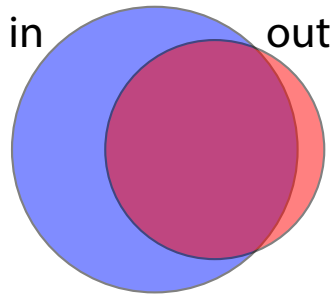
- Comportement spécifique observé sur certains sites de réseautage social comme Twitter
- But : obtenir rapidement un maximum de visibilité
- Qui ils sont : spammers et... célébrités
- Pourquoi les étudier ?
  - Comprendre leur influence sur le réseau
  - Améliorer la qualité de service
  - Appliquer 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

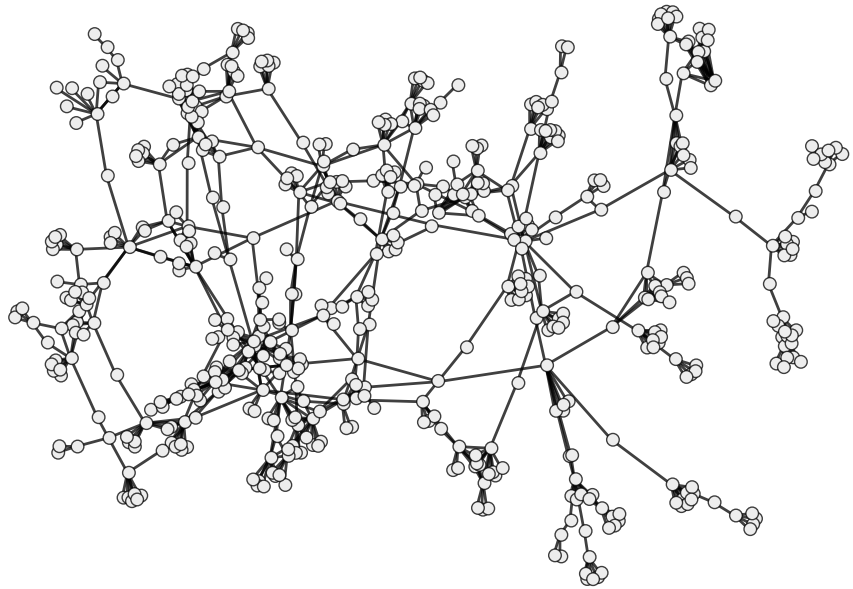






Méthode de Guimerà & Amaral [GA05]

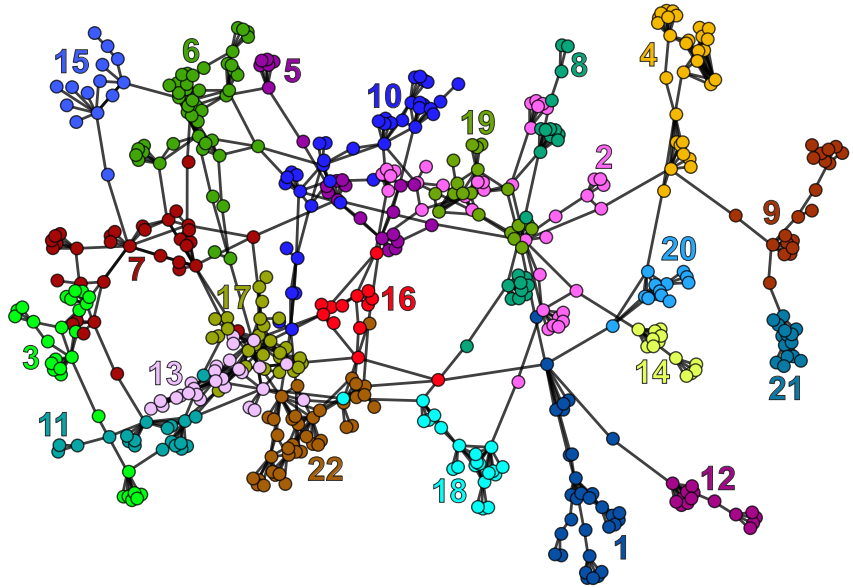
- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :



[LB12]

Méthode de Guimerà & Amaral [GA05]

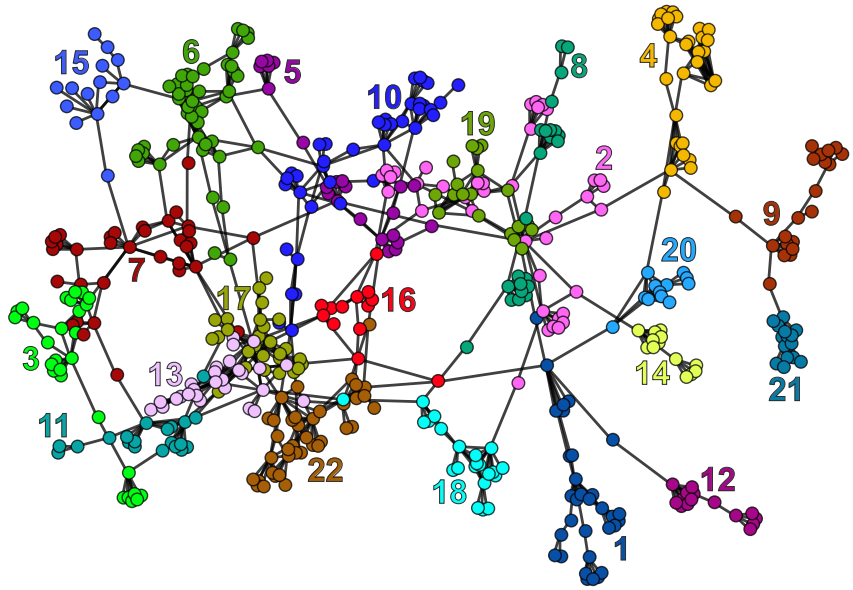
- **Principe** :
  - Caractériser la position d'un nœud en fonction de sa connectivité communautaire
  - Connectivité communautaire décrite par 2 mesures
- **Processus** :
  1. Identification des communautés



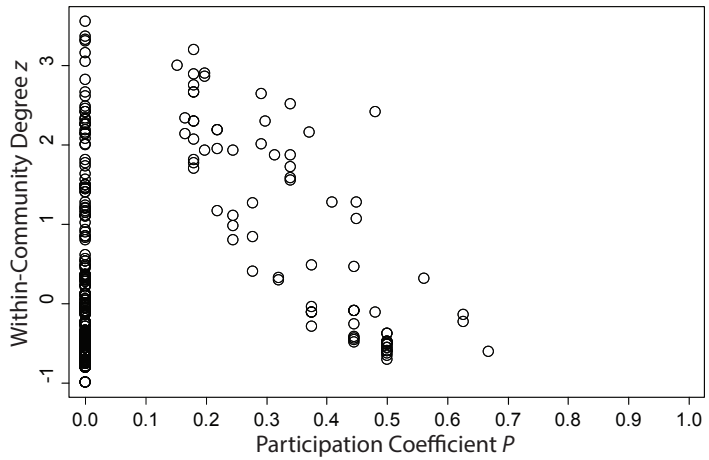
[LB12]

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

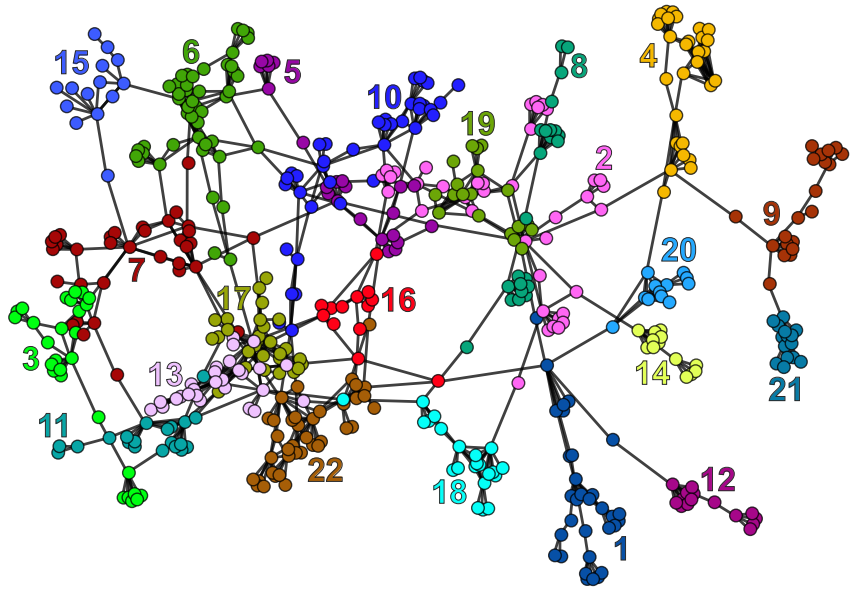


[LB12]

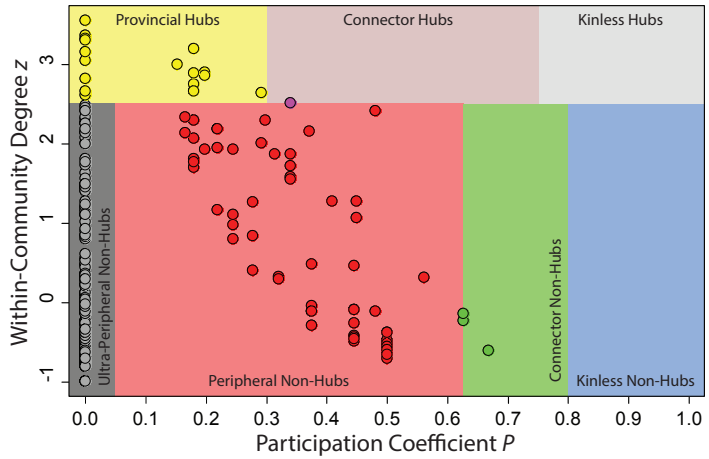


Méthode de Guimerà & Amaral [GA05]

- **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

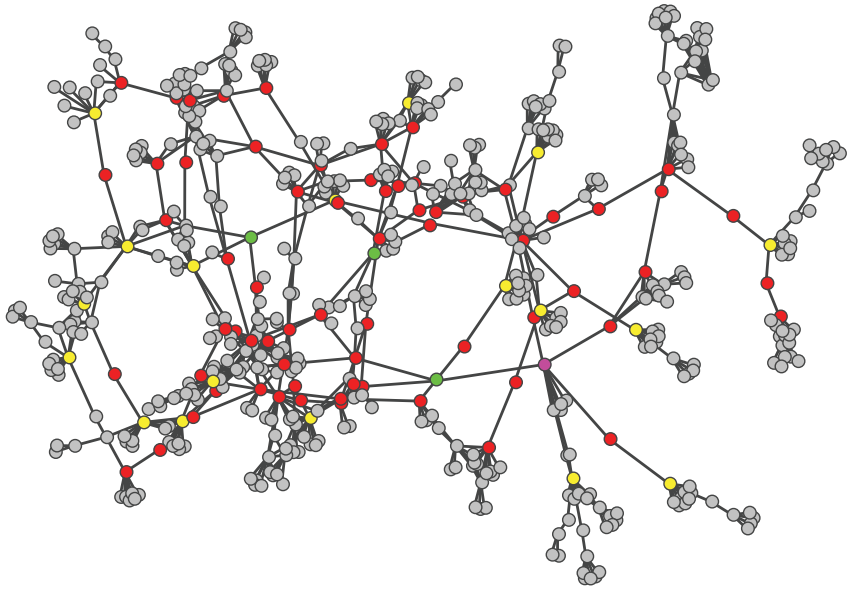


[LB12]

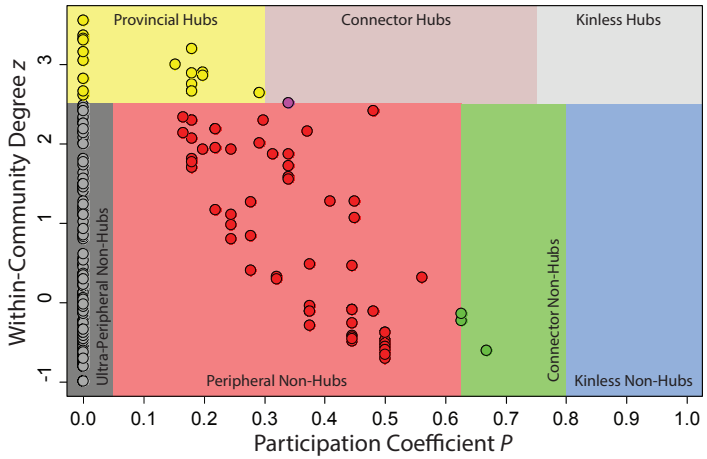


Méthode de Guimerà & Amaral [GA05]

- **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]

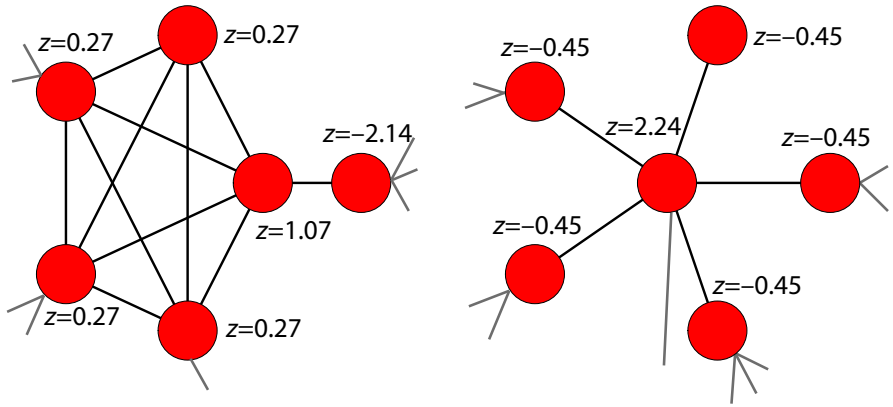


Mesures de rôle

- **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

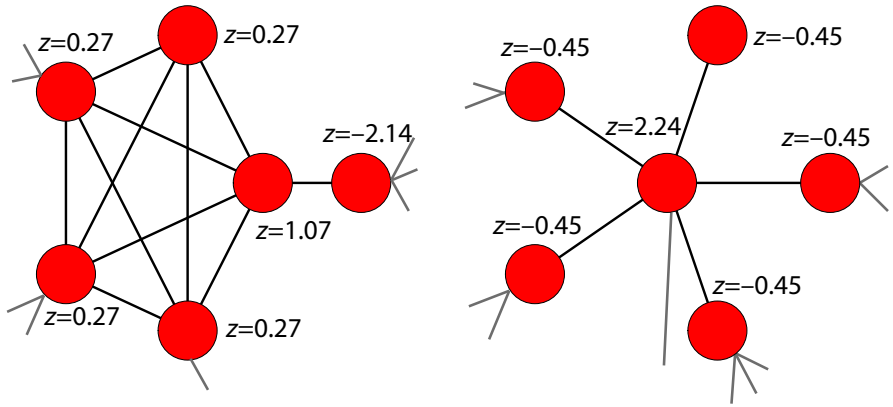
Mesures de rôle

- **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



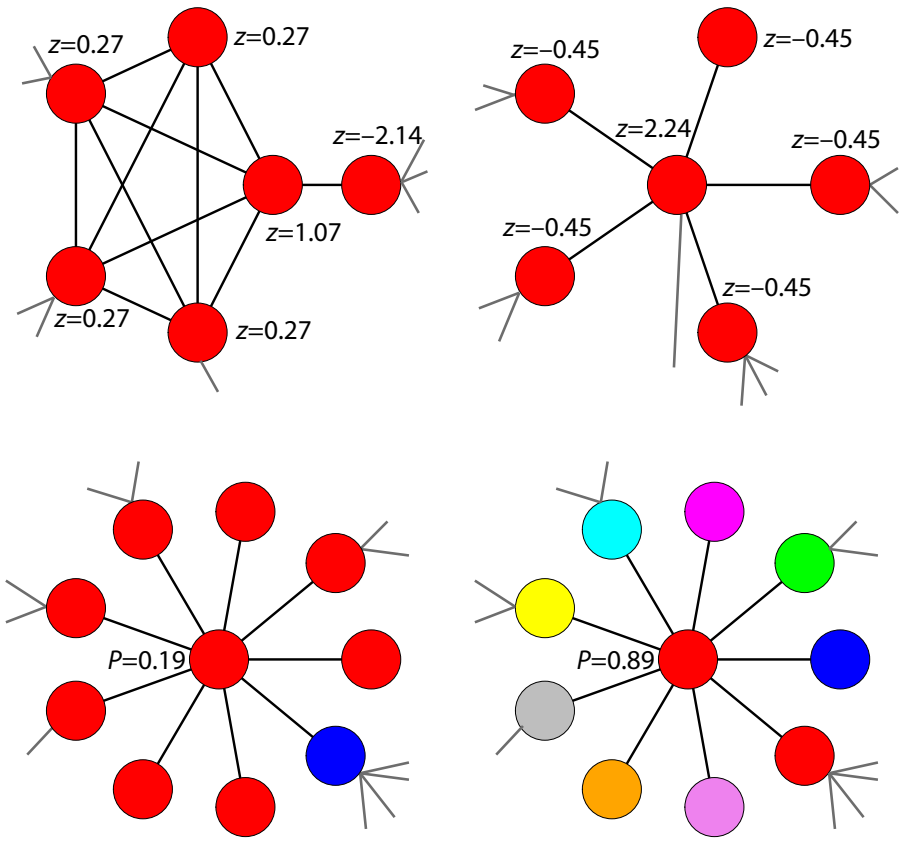
Mesures de rôle

- **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



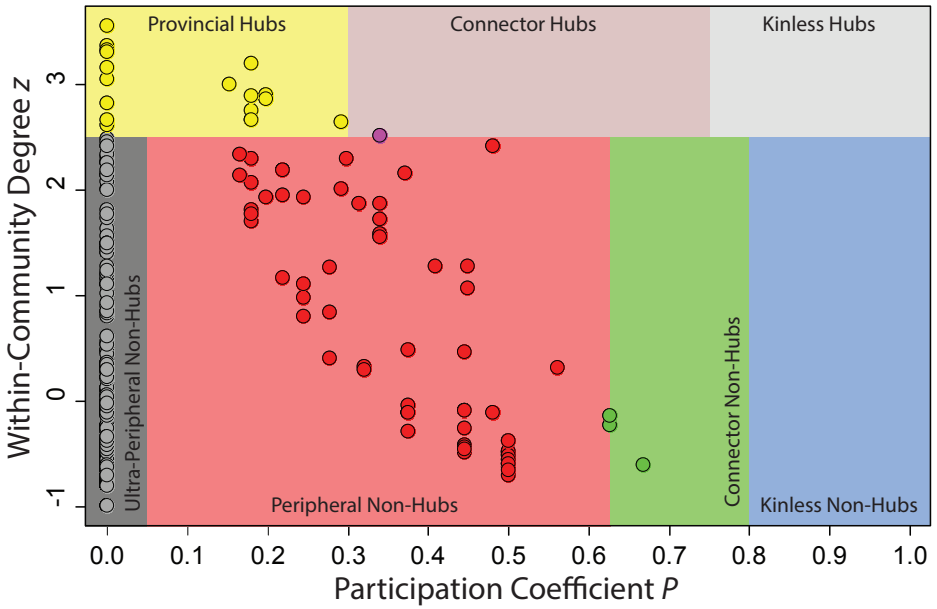
Mesures de rôle

- **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



Limitations de l'approche

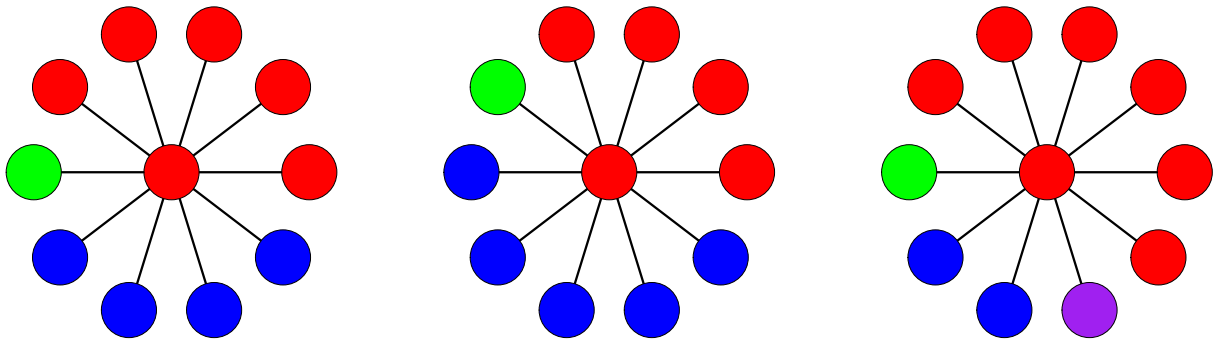
- Orientation des liens ignorée
  - Systèmes à relations asymétriques
  - Twitter : suiveur / suivi

Limitations de l'approche

- Orientation des liens ignorée
  - Systèmes à relations asymétriques
  - Twitter : suiveur / suivi
- Hypothèse d'universalité des seuils
  - Amplitude de z pas normalisée
  - $\rightarrow$ pertinence sur d'autres données ?

Limitations de l'approche

- Orientation des liens ignorée
  - Systèmes à relations asymétriques
  - Twitter : suiveur / suivi
- Hypothèse d'universalité des seuils
  - Amplitude de z pas normalisée
  - $\rightarrow$ pertinence sur d'autres données ?
- Imprécision du coefficient de participation
  - Degré, nombre de communautés, distribution des liens
  - Liens externes, mais aussi internes



$P = 0.58$

Connectivité externe

- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ

Connectivité externe

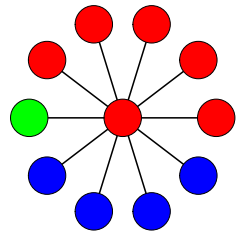
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}

Connectivité externe

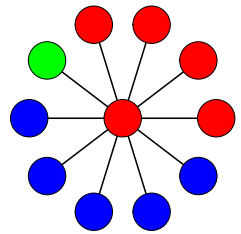
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité**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

Connectivité externe

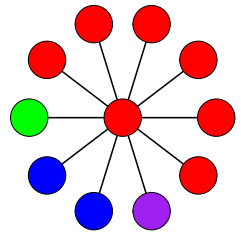
- Restriction aux communautés externes
- 3 aspects distincts considérés :
  - **Diversité** D
    - $\epsilon(u)$: nombre de communautés externes
    - $D(u)$: z-score de ϵ
  - **Intensité externe** I_{ext}
    - k_{ext} : nombre de liens externes
    - $I_{ext}(u)$: z-score de k_{ext}
  - **Hétérogénéité**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



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



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



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

Connectivité interne

- Connectivité interne : mesure z
  - Renommée *intensité interne*
  - Notée I_{int}
- Orientation des liens :
  - Chaque mesure existe en versions entrante et sortante
  - $\rightarrow$ Total de 8 mesures

Identification non-supervisée des rôles

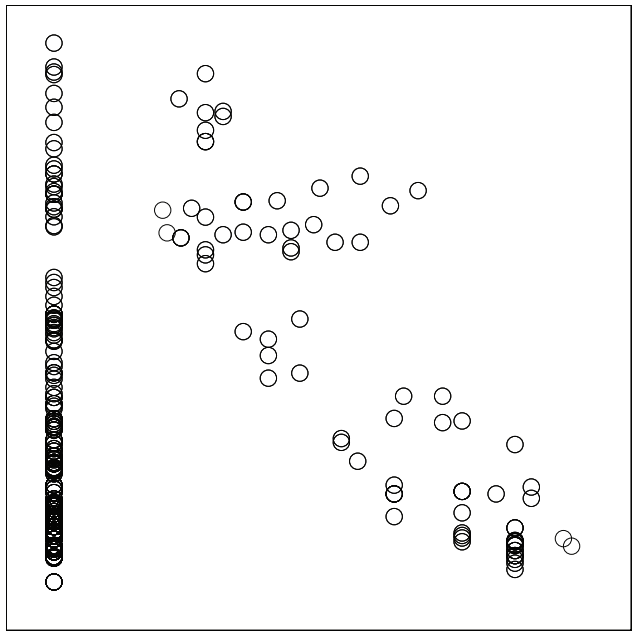
- Difficultés à traiter
  - Traitement des nombreuses mesures
  - Mesures sans bornes fixées
  - Variabilité des données
  - Certains rôles non remplis

Identification non-supervisée des rôles

- Difficultés à traiter
  - Traitement des nombreuses mesures
  - Mesures sans bornes fixées
  - Variabilité des données
  - Certains rôles non remplis
- Analyse de regroupement (clustering)

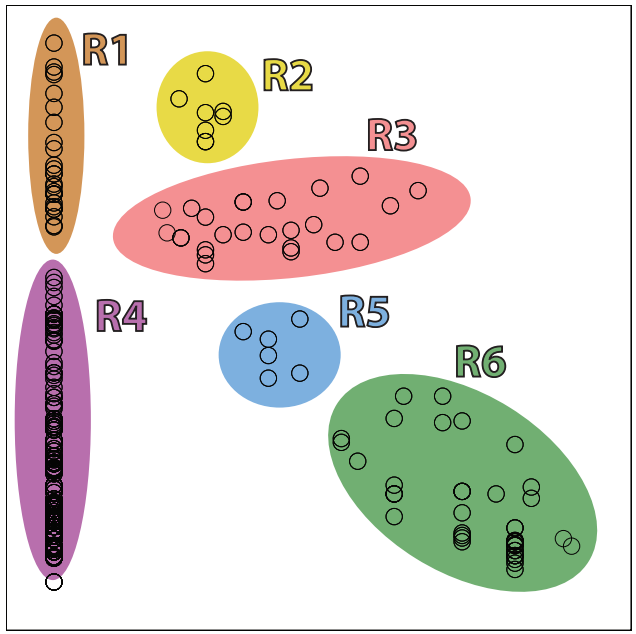
Identification non-supervisée des rôles

- Difficultés à traiter
  - Traitement des nombreuses mesures
  - Mesures sans bornes fixées
  - Variabilité des données
  - Certains rôles non remplis
- Analyse de regroupement (clustering)
  - Appliquée à toutes les mesures simultanément



Identification non-supervisée des rôles

- Difficultés à traiter
  - Traitement des nombreuses mesures
  - Mesures sans bornes fixées
  - Variabilité des données
  - Certains rôles non remplis
- Analyse de regroupement (clustering)
  - Appliquée à toutes les mesures simultanément
  - Chaque groupe obtenu correspond à un rôle



Données & outils

- Réseau étudié
  - Collecté en 2009 [CHBG10]
  - 55 millions de nœuds (utilisateurs)
  - 2 milliards de liens orientés (followee $\rightarrow$ follower)
- Outils
  - Détection de communautés : Louvain
  - Analyse de regroupement : k-moyennes distribué [Lia09]
  - Sélection des groupes : indice Davies-Bouldin [DB79]
  - Code source : `https://github.com/CompNet/Orleans`

Propriétés des groupes

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

Taille des groupes

Propriétés des groupes

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

Taille des groupes

G	I_{int}		D		I_{ext}		H	
1	−0,12	−0,03	−0,55	−0,80	−0,09	−0,04	−0,12	−0,06
2	94,22	311,27	7,18	88,40	113,87	283,79	112,79	285,57
3	5,52	1,40	5,60	3,10	5,28	1,43	6,76	2,34
4	−0,04	0,00	−0,37	0,69	−0,07	0,00	−0,10	−0,01
5	−0,03	−0,01	0,60	0,19	−0,03	−0,02	−0,04	−0,02
6	0,48	0,12	1,96	1,70	0,35	0,12	0,53	0,19

Valeurs des mesures par groupe

Types de capitalistes sociaux

- Capitalistes sociaux identifiés avec la méthode de [DP13]
  - Méthode basée sur 2 mesures topologiques : *ratio* et *indice de chevauchement*
  - 200000 utilisateurs détectés ($0,4\%$ des nœuds)
- Types de capitalistes sociaux
  - Degré total k : efficacité d'exécution
    - Degré faible ($500 < k < 10000$) : 193300 nœuds
    - Degré élevé ($k \geq 10000$) : 6700 nœuds
  - Ratio r : nombre de followees divisé par nombre de followers
    - Degré faible : FMIFY ($r < 1$) ou IFYFM ($r \geq 1$)
    - Degré élevé : passifs ($r < 0,7$), FMIFY ($0,7 \leq r < 1$) ou IFYFM ($r > 1$)

Répartition dans les groupes

Ratio	**G1** NU	**G2** PO	**G3** PC	**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

Répartition dans les groupes

Ratio	**G1** NU	**G2** PO	**G3** PC	**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
  - Constituent 38% de G2 et 14% de G3
  - Forte concentration dans G6 (60% des faibles degrés) et G3 (90% des degrés élevés)
  - Pivot : G2 et G3
  - Connecteur & orphelin : G2, G3 et G6
- Niveau stratégique
  - Degré : discriminant
  - Ratio : pas de différence importante entre FMIFY et IFYFM
  - Confirmation mode passif : 31% de G2 (mesures entrantes)
  - Diversité sortante élevée (G3, G6) $\rightarrow$ suivre de nombreuses communautés

Conclusion

- Contributions
  - Extension orientée
  - Mesures supplémentaires pour la connectivité externe
  - Méthode non-supervisée pour déterminer les rôles
  - Analyse des capitalistes sociaux dans Twitter
- Perspectives
  - Application à d'autres systèmes
  - Prise en compte des poids
  - Communautés recouvrantes

Références I

[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.

[DB79] David Davies and Donald Bouldin.

A cluster separation measure.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2) :224–227, 1979.

[DP13] Nicolas Dugué and Anthony Perez.

Detecting social capitalists on twitter using similarity measures.

In *Complex Networks IV*, volume 476 of *Studies in Computational Intelligence*, pages 1–12. Springer, 2013.

Références II

[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 III

[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.