Identifying the Community Roles of Social Capitalists in the Twitter Network

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

[1]University of Orléans, LIFO
{nicolas.dugue,anthony.perez}@univ-orleans.fr

[2]Galatasaray University, Computer Science Department
vlabatut@gsu.edu.tr

ASONAM
20 August 2014

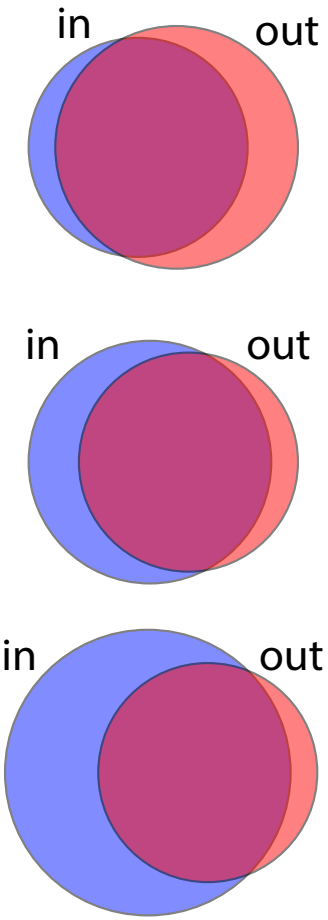
Outline

Notion of Social Capitalism [GVK+12]

- Specific behavior observed on certain social networking websites such as Twitter
- Goal : quickly obtain a maximal visibility
- Who they are : spammers and... celebrities
- Why study them ?
  - Understand their influence on the network
  - Improve quality of service
  - Apply their methods to other domains (e.g. marketing)

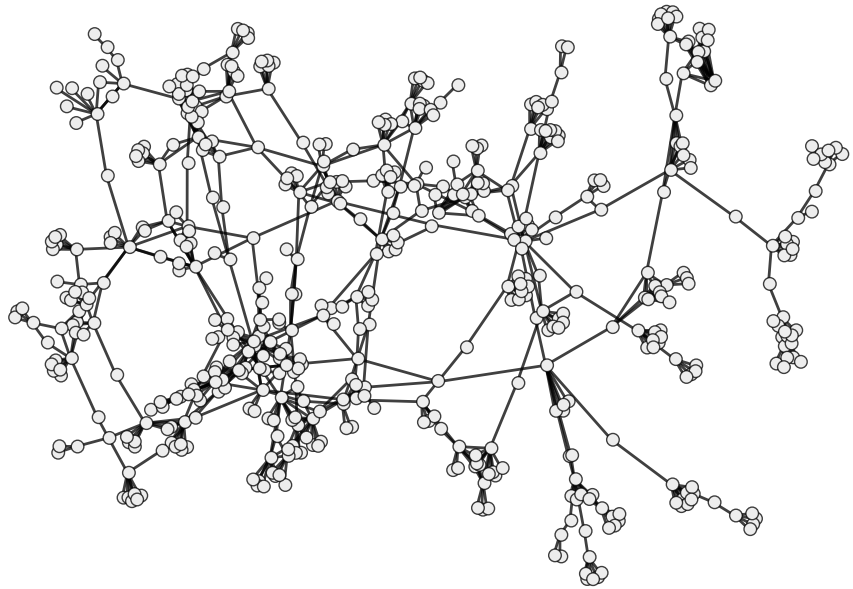
Strategies of Social Capitalists

- I Follow You, Follow Me (IFYFM)
- Follow Me, I Follow You (FMIFY)
- Passive state



Method of Guimerà & Amaral [GA05]

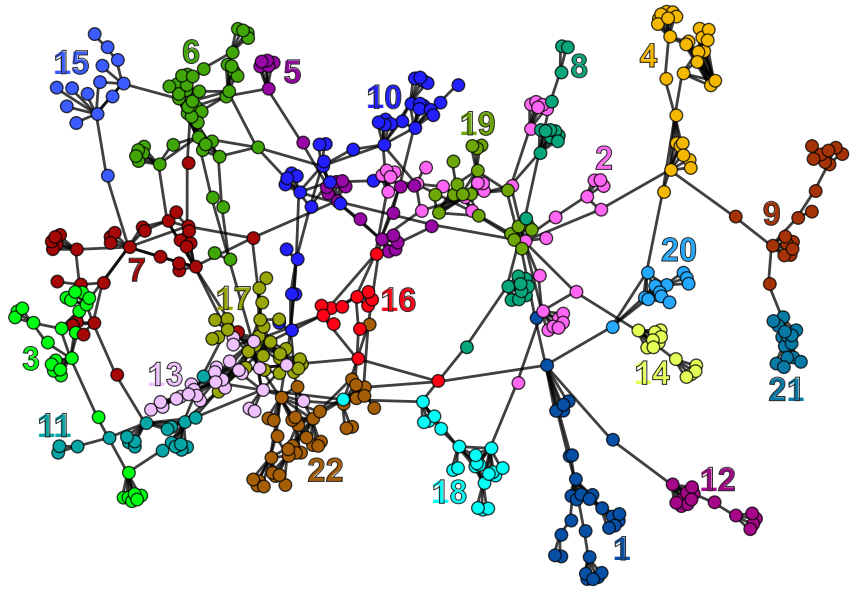
- **Principle** :
  - Position of a node expressed in terms of its *community connectivity*
  - Community connectivity described by 2 measures
- **Process** :



[LB12]

Method of Guimerà & Amaral [GA05]

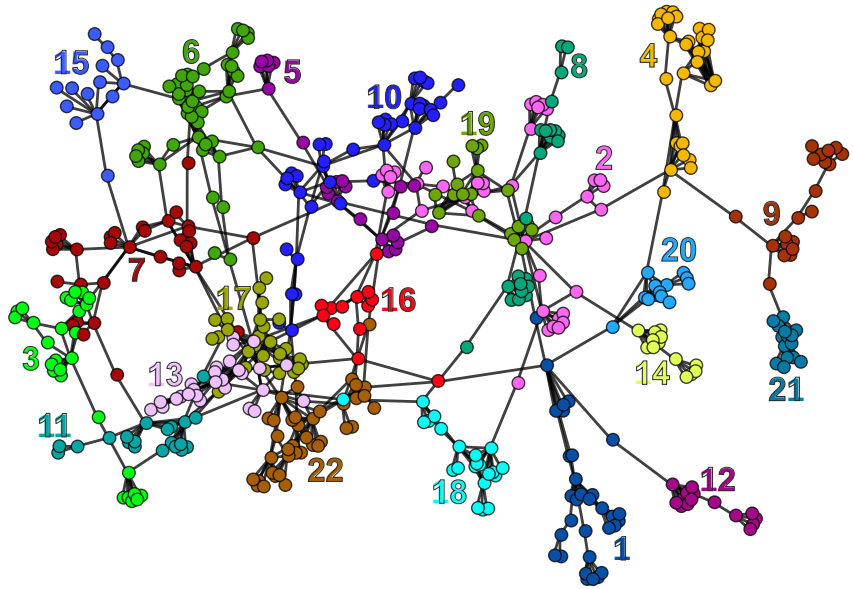
- **Principle** :
  - Position of a node expressed in terms of its *community connectivity*
  - Community connectivity described by 2 measures
- **Process** :
  1. Identify communities



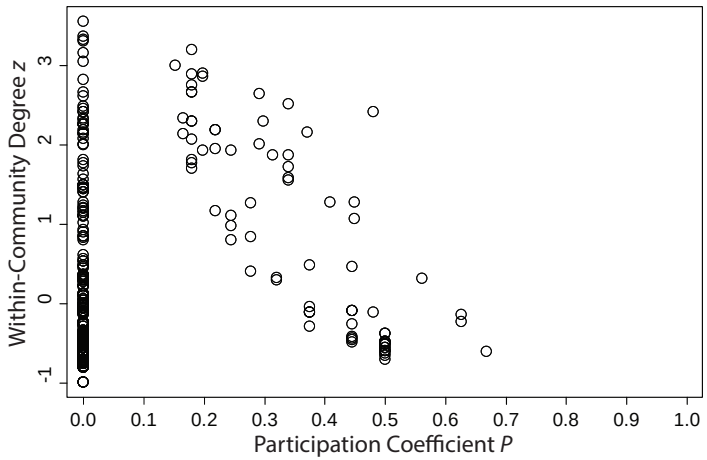
[LB12]

Method of Guimerà & Amaral [GA05]

- **Principle** :
  - Position of a node expressed in terms of its *community connectivity*
  - Community connectivity described by 2 measures
- **Process** :
  1. Identify communities
  2. Process both nodal measures

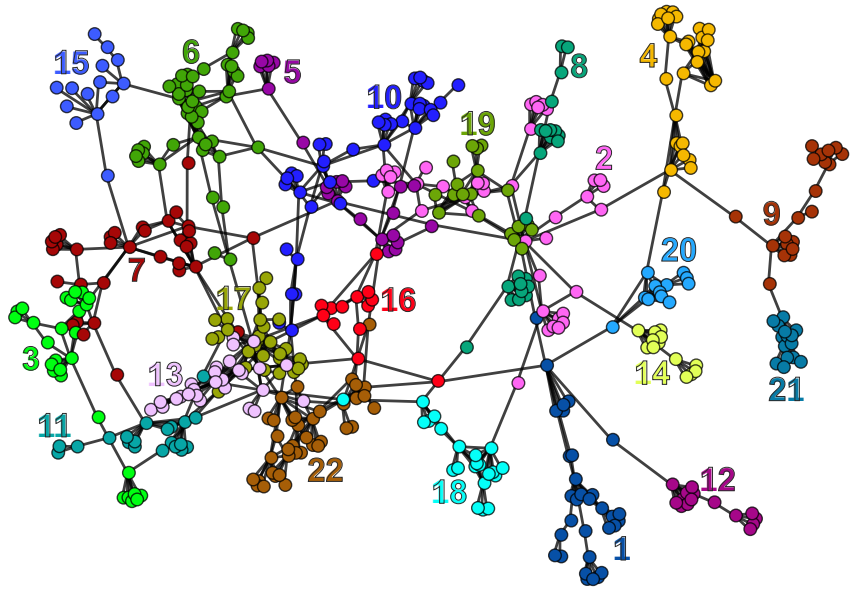


[LB12]

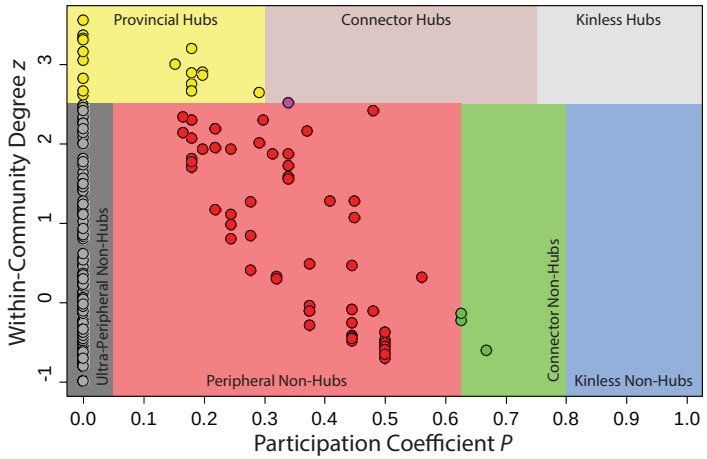


Method of Guimerà & Amaral [GA05]

- **Principle** :
  - Position of a node expressed in terms of its *community connectivity*
  - Community connectivity described by 2 measures
- **Process** :
  1. Identify communities
  2. Process both nodal measures
  3. Partition the resulting $2D$-space

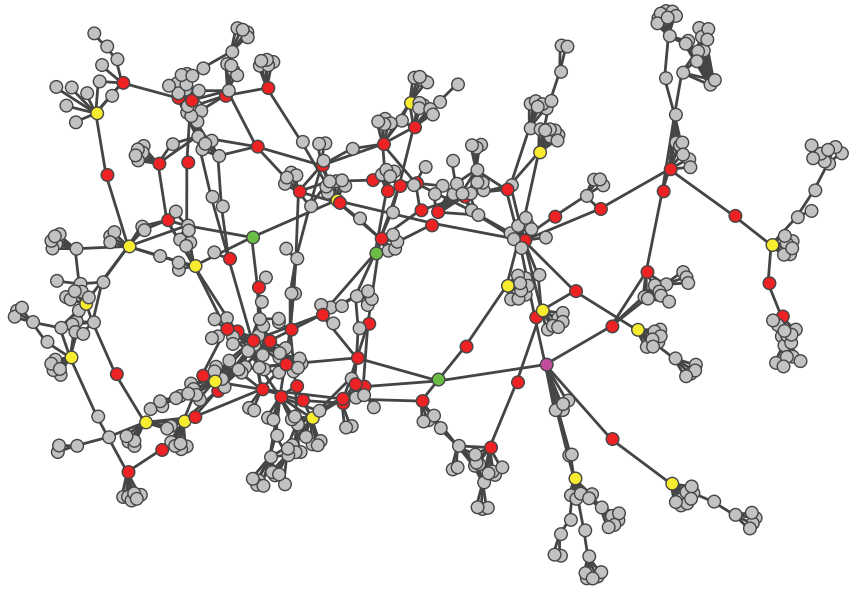


[LB12]

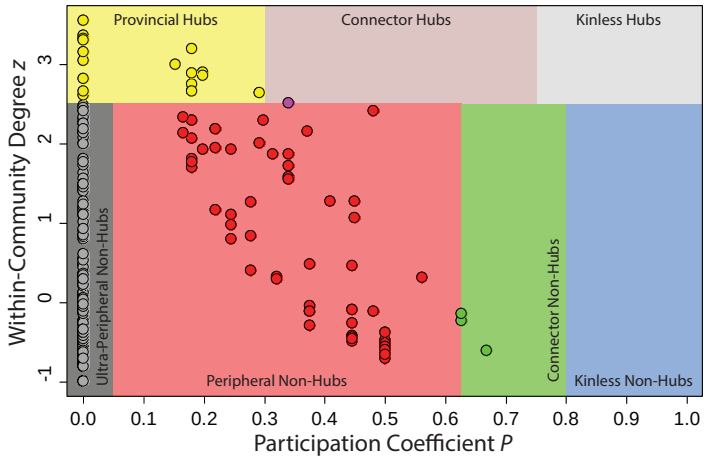


Method of Guimerà & Amaral [GA05]

- **Principle** :
  - Position of a node expressed in terms of its *community connectivity*
  - Community connectivity described by 2 measures
- **Process** :
  1. Identify communities
  2. Process both nodal measures
  3. Partition the resulting $2D$-space
  4. Match roles to the obtained parts



[LB12]

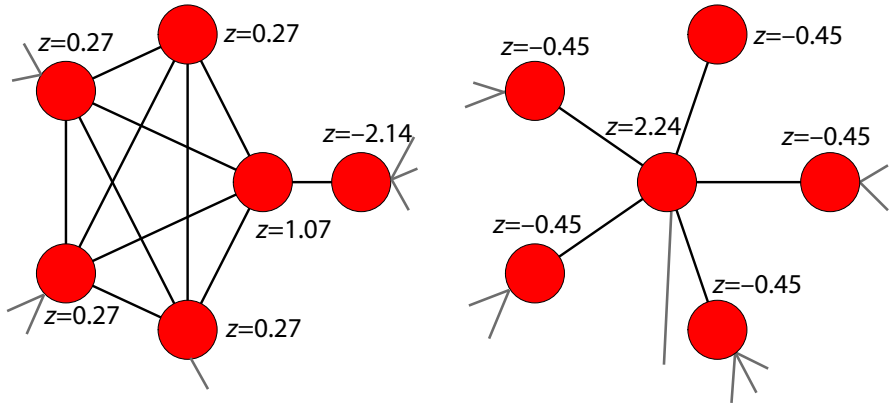


Role Measures

- **Normalized Internal Degree**
  - *Internal* connectivity
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score of the internal degree k_{int}
  - No fixed bounds

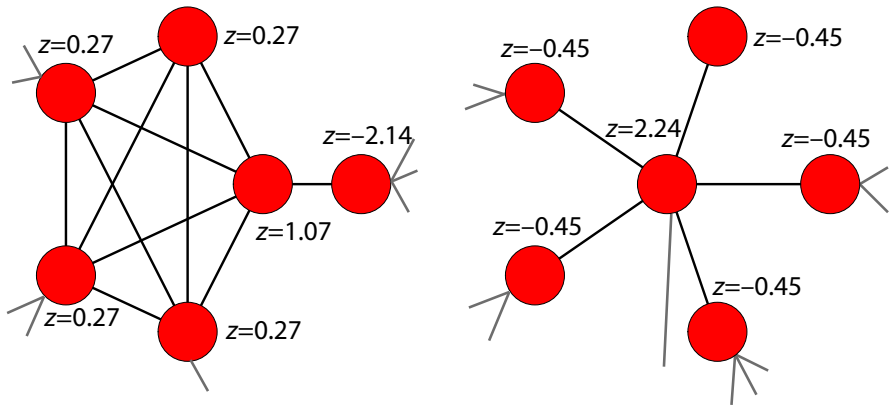
Role Measures

- **Normalized Internal Degree**
  - *Internal* connectivity
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score of the internal degree k_{int}
  - No fixed bounds



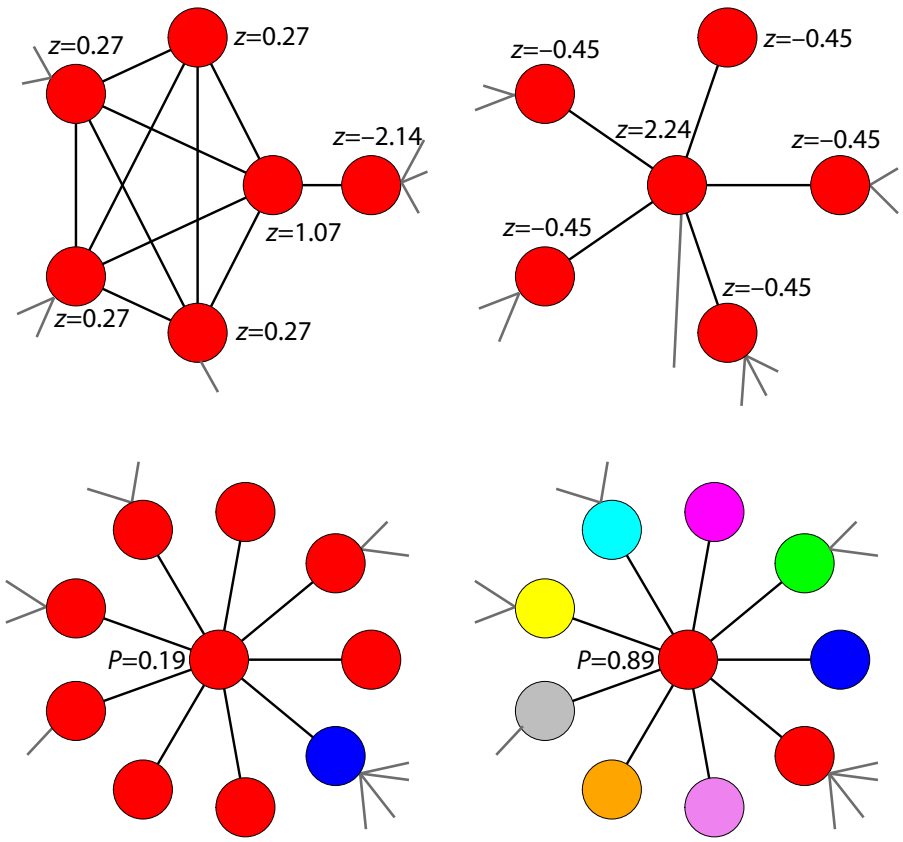
Role Measures

- **Normalized Internal Degree**
  - *Internal* connectivity
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score of the internal degree k_{int}
  - No fixed bounds
- **Participation Coefficient**
  - *External* connectivity
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)} \right)^2$
  - k_i : degree for C_i
  - $P(u) = 0$:
    - Single community
  - $P(u) \approx 1$:
    - Many communities
    - Same number of links



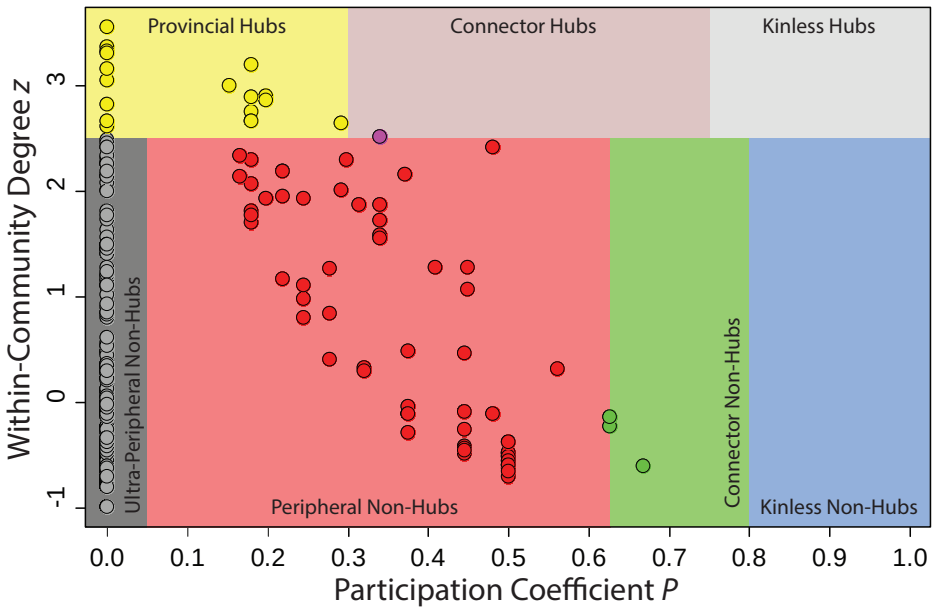
Role Measures

- **Normalized Internal Degree**
  - *Internal* connectivity
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score of the internal degree k_{int}
  - No fixed bounds
- **Participation Coefficient**
  - *External* connectivity
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)}\right)^2$
  - k_i : degree for C_i
  - $P(u) = 0$:
    - Single community
  - $P(u) \approx 1$:
    - Many communities
    - Same number of links



Role Measures

- **Normalized Internal Degree**
  - *Internal* connectivity
    $z(u) = \frac{k_{int}(u) - \mu_i(k_{int})}{\sigma_i(k_{int})}$, $u \in C_i$
  - z-score of the internal degree k_{int}
  - No fixed bounds
- **Participation Coefficient**
  - *External* connectivity
    $P(u) = 1 - \sum_i \left(\frac{k_i(u)}{k(u)}\right)^2$
  - k_i : degree for C_i
  - $P(u) = 0$:
    - Single community
  - $P(u) \approx 1$:
    - Many communities
    - Same number of links



Limitations of the Approach

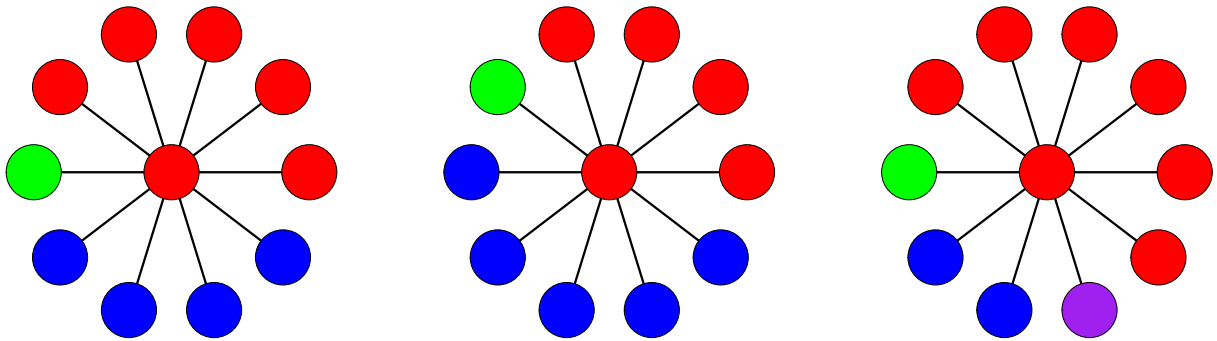
- Link directions ignored
  - Systems with asymmetric relationships
  - Twitter : follower vs. followee

Limitations of the Approach

- Link directions ignored
  - Systems with asymmetric relationships
  - Twitter : follower vs. followee
- Universal threshold assumption
  - Amplitude of z not normalized
  - $\rightarrow$ relevance on other data ?

Limitations of the Approach

- Link directions ignored
  - Systems with asymmetric relationships
  - Twitter : follower vs. followee
- Universal threshold assumption
  - Amplitude of z not normalized
  - $\rightarrow$ relevance on other data ?
- Imprecision of the participation coefficient
  - Degree, number of communities, link distribution
  - External, but also internal links



$P = 0.58$

External Connectivity

- Restriction to external communities only
- 3 distinct aspects are considered :
  - **Diversity** D
    - $\epsilon(u)$: number of external communities
    - $D(u)$: z-score of ϵ

External Connectivity

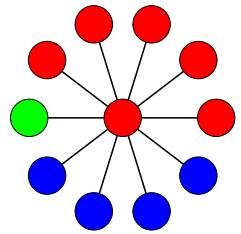
- Restriction to external communities only
- 3 distinct aspects are considered :
  - **Diversity** D
    - $\epsilon(u)$: number of external communities
    - $D(u)$: z-score of ϵ
  - **External Intensity** I_{ext}
    - k_{ext} : number of external links
    - $I_{ext}(u)$: z-score of k_{ext}

External Connectivity

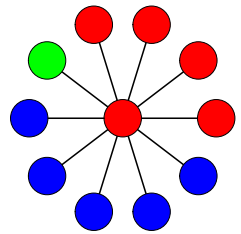
- Restriction to external communities only
- 3 distinct aspects are considered :
  - **Diversity** D
    - $\epsilon(u)$: number of external communities
    - $D(u)$: z-score of ϵ
  - **External Intensity** I_{ext}
    - k_{ext} : number of external links
    - $I_{ext}(u)$: z-score of k_{ext}
  - **Heterogeneity** H
    - Dispersion of the external links
    - $\lambda(u)$: standard deviation of k_i
    - $H(u)$: z-score of λ

External Connectivity

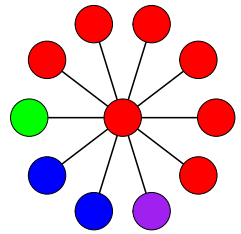
- Restriction to external communities only
- 3 distinct aspects are considered :
  - **Diversity** D
    - $\epsilon(u)$: number of external communities
    - $D(u)$: z-score of ϵ
  - **External Intensity** I_{ext}
    - k_{ext} : number of external links
    - $I_{ext}(u)$: z-score of k_{ext}
  - **Heterogeneity** H
    - Dispersion of the external links
    - $\lambda(u)$: standard deviation of k_i
    - $H(u)$: z-score of λ



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



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



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

Internal Connectivity

- Internal connectivity : measure z
  - Renamed *Internal Intensity*
  - Noted I_{int}
- Link directions :
  - Each measure is derived in two versions :
    - Incoming links
    - Outgoing links
  - $\rightarrow$ Total of 8 measures

Unsupervised Identification of the Roles

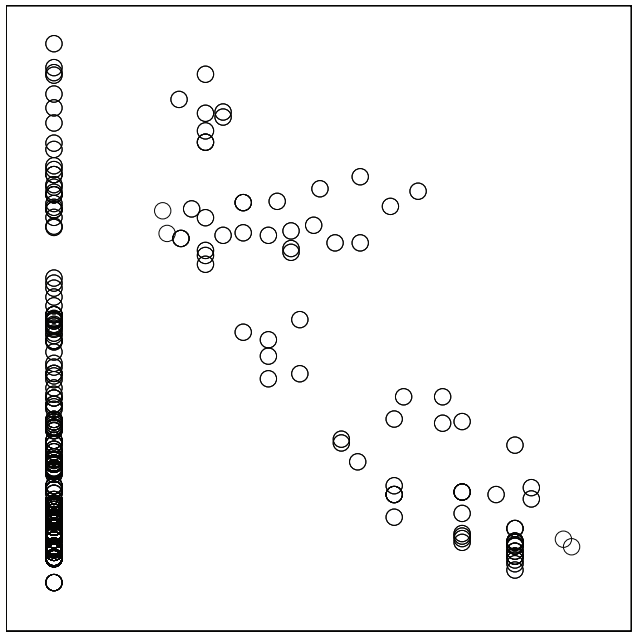
- Several difficulties must be dealt with :
  - Processing many measures
  - Measures without fixed bounds
  - Variability of the data
  - Some roles possibly not fulfilled

Unsupervised Identification of the Roles

- Several difficulties must be dealt with :
  - Processing many measures
  - Measures without fixed bounds
  - Variability of the data
  - Some roles possibly not fulfilled
- How : Cluster Analysis

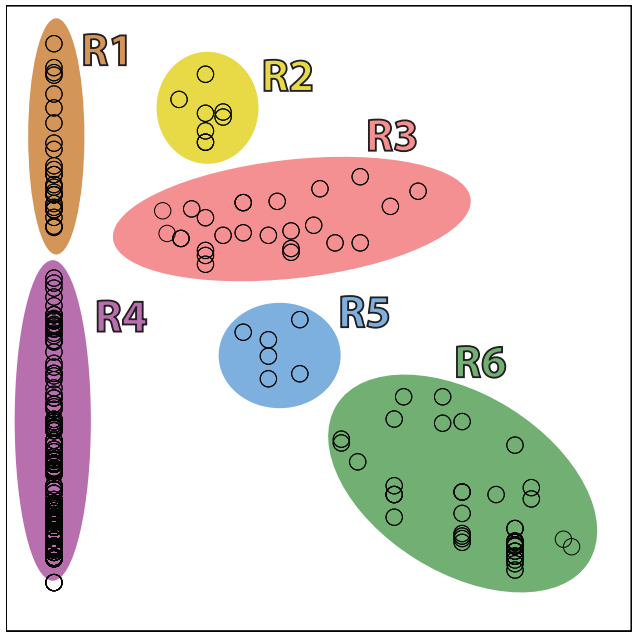
Unsupervised Identification of the Roles

- Several difficulties must be dealt with :
  - Processing many measures
  - Measures without fixed bounds
  - Variability of the data
  - Some roles possibly not fulfilled
- How : Cluster Analysis
  - Applied to all 8 measures simultaneously



Unsupervised Identification of the Roles

- Several difficulties must be dealt with :
  - Processing many measures
  - Measures without fixed bounds
  - Variability of the data
  - Some roles possibly not fulfilled
- How : Cluster Analysis
  - Applied to all 8 measures simultaneously
  - Each detected cluster corresponds to a role



Data & Tools

- Studied network
  - Collected in 2009 [CHBG10]
  - 55 million nodes (users)
  - 2 billion directed links (followee $\rightarrow$ follower)
- Tools
  - Community detection : Louvain [BGLL08]
  - Cluster analysis : distributed k-means [Lia09]
  - Cluster quality : Davies-Bouldin index [DB79]
  - Source code : `https://github.com/CompNet/Orleans`

Cluster Properties

Cluster	Size	Proportion	Role
1	$24,543,667$	46.68%	Ultra-peripheral non-hubs
2	304	$< 0.01\%$	Kinless hubs
3	$303,674$	0.58%	Connector hubs
4	$11,929,722$	22.69%	Incoming peripheral non-hubs
5	$10,828,599$	20.59%	Outgoing peripheral non-hubs
6	$4,973,717$	9.46%	Connector non-hubs

Clusters and their corresponding roles

Cluster Properties

Cluster	Size	Proportion	Role
1	24,543,667	46.68%	Ultra-peripheral non-hubs
2	304	$< 0.01\%$	Kinless hubs
3	303,674	0.58%	Connector hubs
4	11,929,722	22.69%	Incoming peripheral non-hubs
5	10,828,599	20.59%	Outgoing peripheral non-hubs
6	4,973,717	9.46%	Connector non-hubs

Clusters and their corresponding roles

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

Measures values for all clusters

Types of Social Capitalists

- Social capitalists identified with Dugué & Perez's method [DP13]
  - Based on 2 topological measures : *ratio* and *overlap index*
  - $200,000$ users considered as social capitalists in our data (0.4% of all nodes)
- Types of social capitalists
  - Total degree k :
    - Reflects success (in becoming a social capitalist)
    - Low degree ($500 < k < 10,000$) : $193,300$ nodes
    - High degree ($k \geq 10,000$) : $6,700$ nodes
  - Ratio r : number of followees divided by number of followers
    - Indicates the selected strategy
    - Low degree : FMIFY ($r < 1$) or IFYFM ($r \geq 1$)
    - High degree : passive ($r < 0,7$), FMIFY ($0,7 \leq r < 1$) or IFYFM ($r > 1$)

Distribution Over Clusters

Ratio	**C1** UN	**C2** KH	**C3** CH	**C4** PN(I)	**C5** PN(O)	**C6** CN
$r < 1$	0.01%	0.00%	**23.10%**	3.42%	**18.28%**	**55.19%**
	$< 0.01\%$	0.00%	3.71%	0.14%	0.08%	0.54%
$r > 1$	0.03%	0.00%	**18.78%**	0.48%	**14.31%**	**66.40%**
	$< 0.01\%$	0.00%	**6.61%**	$< 0.01\%$	0.14%	1.43%

Low degree capitalists ($500 < k < 10,000$)

Distribution Over Clusters

Ratio	**C1** UN	**C2** KH	**C3** CH	**C4** PN(I)	**C5** PN(O)	**C6** CN
$r < 1$	0.01%	0.00%	**23.10%**	3.42%	**18.28%**	**55.19%**
	$< 0.01\%$	0.00%	3.71%	0.14%	0.08%	0.54%
$r > 1$	0.03%	0.00%	**18.78%**	0.48%	**14.31%**	**66.40%**
	$< 0.01\%$	0.00%	**6.61%**	$< 0.01\%$	0.14%	1.43%

Low degree capitalists ($500 < k < 10,000$)

Ratio	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
$r < 0.7$	0.00%	**12.14%**	**87.29%**	0.00%	0.00%	0.57%
	0.00%	**21.05%**	0.15%	0.00%	0.00%	$< 0.01\%$
$0.7 < r < 1$	0.00%	1.55%	**95.64%**	0.00%	0.00%	2.81%
	0.00%	**7.24%**	0.45%	0.00%	0.00%	$< 0.01\%$
$r > 1$	0.00%	0.03%	**97.99%**	0.00%	0.00%	1.98%
	0.00%	0.33%	1.22%	0.00%	0.00%	$< 0.01\%$

High degree capitalists ($k \geq 10,000$)

Summary of Observations

- Social capitalists hold only certain roles
  - 3 clusters for low degrees : #3, #5 and #6
  - 2 clusters for high degrees : #2 and #3
- Role depends more on success than strategy itself
  - Clusters #5 and #6 : beginners
  - Cluster #3 : intermediary
  - Cluster #2 : confirmed

Conclusion

- Contributions
  - Directed extension of Guimerá & Amaral's measures
  - Additional measures for external connectivity
  - Unsupervised method to determine roles
  - Analysis of social capitalists in Twitter
- Perspectives
  - Application to other systems than Twitter
  - Take weighted links into account
  - Deal with overlapping communities

References I

[BGLL08] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre.

Fast unfolding of communities in large networks.

J. Stat. Mech., 10 :P10008, Oct 2008.

[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.

References II

[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.

[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.

References 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.