Identifying the Community Roles of Social Capitalists in the Twitter Network

Nicolas Dugué, Vincent Labatut, Anthony Perez

▶ **To cite this version:**

Nicolas Dugué, Vincent Labatut, Anthony Perez. Identifying the Community Roles of Social Capitalists in the Twitter Network. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM), Aug 2014, Pékin, China. pp.371-374, 10.1109/ASONAM.2014.6921612 . hal-01011910v2

HAL Id: hal-01011910
https://hal.science/hal-01011910v2

Submitted on 25 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Identifying the Community Roles of Social Capitalists in the Twitter Network

Vincent Labatut
Galatasaray University, Computer Science Department
Çırağan cad. n36, Ortaköy 34357,
İstanbul, Turquie

Nicolas Dugué, Anthony Perez
Univ. Orléans, INSA Centre Val de Loire
LIFO EA 4022 45067 Orléans, France

***Abstract*—In the context of Twitter, social capitalists are specific users trying to increase their number of followers and interactions by any means. These users are not healthy for the Twitter network since they flaw notions of influence and visibility. Indeed, it has recently been observed that they are real and active users that can help malicious users such as spammers gaining influence. Studying their behavior and understanding their position in Twitter is thus of important interest. A recent work provided an efficient way to detect social capitalists using two simple topological measures. Based on this detection method, we study how social capitalists are distributed over Twitter's friend-to-follower network. We are especially interested in analyzing how they are organized, and how their links spread across the network. Answering these questions allows to know whether the social capitalism methods increase the actual visibility on the service. To that aim, we study the position of social capitalists on Twitter w.r.t. the community structure of the network. We base our work on the concept of community role of a node, which describes its position in a network depending on its connectivity at the community level. The topological measures originally defined to characterize these roles consider only some aspects of community-related connectivity and rely on a set of empirically fixed thresholds. We first show the limitations of such measures and then extend and generalize them by considering new aspects of the community-related connectivity. Moreover, we use an unsupervised approach to distinguish the roles, in order to provide more flexibility relatively to the studied system. We then apply our method to the case of social capitalists and show that they are highly visible on Twitter, due to the specific roles they occupy.**

I. Introduction

Context. The last decade has been marked by an increase in both the number of online social networking services and the number of users of such services. This observation is particularly relevant when considering Twitter, which had 200 millions accounts in April 2011 [1] and reached 500 millions accounts in October 2012 [2]. Twitter is mostly used to share, seek and debate about information, or to let the world know about daily events [3]. The amount of information shared on Twitter is considerable: there are about 1 billion tweets posted every two and a half days [4]. While focusing on microblogging, Twitter can be considered as a social networking service, since it includes social features. Indeed, to see the messages of other users, a Twitter user has to *follow* them (i.e. make a subscription). Furthermore, a user can *retweet* [5] other users' tweets, for instance when he finds them interesting and wants to share them with their followers. Besides, users can *mention* other users to draw their attention by adding `@UserName` in their message. Some Twitter users are trying to use these particular properties to spread efficiently some information [6]. One of the simplest way to reach this objective is to gain as many followers as possible, since this gives a higher visibility to the user's tweets when using the network search engines [6].

Social capitalists. These specific users are called *social capitalists*. They have been recently observed and studied by Ghosh *et al.* [6] in a study related to *link-farming* in Twitter. They noticed in particular that users responding the most to the solicitation of spammers are in fact *real*, *active* users. To increase their number of followers, social capitalists use several techniques [6], [7], the most common one being to follow a lot of users *regardless of their content*, just hoping to be followed back. Because of this lack of interest in the content produced by the users they follow, social capitalists are not healthy for a service such as Twitter. Indeed, this behavior helps spammers gaining influence [6], and more generally makes the task of finding relevant information harder for regular users. Studying their behavior and understanding their position in Twitter is therefore a very important task to improve the service, since it can allow designing better search engines or functioning rules. In a recent work, Dugué and Perez [7] have shown that social capitalists can be efficiently detected and classified using two purely topological measures, called *overlap* [8] and *ratio* indices. They provide useful information regarding the interaction between the set of *friends* and the set of *followers*[1] of a user, which are supposed to have a large intersection whenever a user applies social capitalism techniques. In this work, we rely on this detection method to characterize the behavior of social capitalists. To better understand how they are organized, how really visible they are and how their links spread across the network, we study the positions that social capitalists occupy in Twitter w.r.t. the community structure of the network.

Community roles. In its simplest form, the community structure of a complex network can be defined as a partition of its node set, each part corresponding to a community. Community detection methods generally try to perform this partition in order to obtain densely connected groups of nodes, relatively to the rest of the network [9]. Hundreds of such algorithms have been defined in the last ten years, see [10] for a very detailed review of the domain. The notion of community structure is particularly interesting because it allows studying the network at an intermediate level, compared to the more classic global

[1]For a given user, *friends* denote the set of users he follows, and *followers* the set of users that follow him, as per the official Twitter terminology.

(whole network) and local (node neighborhood) approaches.

The concept of community role is a good illustration of this characteristic. It consists in describing a node depending on the position it holds in its own community[2]. Community roles were initially introduced by Guimerà and Amaral [11] to study metabolic networks. After having applied a standard community detection method, they characterize each node according to two *ad hoc* measures, each one describing a specific aspect of the community-related connectivity. The node role is then selected amongst 7 predefined ones by comparing the two values to some empirically fixed thresholds. Guimerà and Amaral [11] showed certain systems possess a role invariance property: when several instances of the system are considered, nodes are different but roles are similarly distributed. Scripps *et al*. [12], apparently unaware of this previous work, later adopted a similar approach, but this time for influence maximization and link-based classification purposes. They also use two measures: first the degree, to assess the intensity of the general node connectivity, and second an *ad hoc* measure, to reflect the number of communities to which it is connected. They then use arbitrary thresholds to define 4 distinct roles.

Our contribution. In this paper, we study the community roles of social capitalists within a freely-available Twitter network provided by Cha *et al*. [13]. We focus on the concept of community role as described by Guimerà and Amaral [11], because it relies more heavily on the community structure. In a first place, we highlight two important limitations of this community role approach. We show that the existing measures do not take into account all aspects of the community-related external connectivity of a node. Moreover, we object the assumption of universality of the thresholds applied to the measures in order to distinguish the different node roles. The dataset we use constitutes a counter-example showing the original thresholds are not relevant for all systems. We then explain how to tackle these limitations. We first introduce three new measures to characterize the external connectivity of a node in a more complete and detailed way. We then describe an unsupervised approach aiming at identifying the node roles without using fixed thresholds. Finally, we apply our method on the Twitter network to determine the position of social capitalists, and show they occupy specific roles in the network. In particular, most of them are well connected to their community, and overall a large part of them spread their links outside their community very efficiently. This gives meaningful insights regarding the actual visibility of these users. They thus seem to occupy roles leading to a high visibility in Twitter.

Outline. We first present the concept of social capitalists in Twitter in more details (Section II). Next, we describe the method proposed by Guimerà and Amaral [11] to identify the community roles of nodes (Section III-A) and provide some elements towards its limitation (Section III-B). We then describe the solutions we propose to tackle these limitations (Section III-C) and finally apply our method to study the roles of social capitalists in Twitter (Section IV).

[2]Note that the notion of role also appears in works related to block modeling, but it is not defined in terms of position in a community [11].

II. Social Capitalists in Twitter

Social capitalists have first been highlighted by Ghosh *et al*. [6] during a study focused on *link-farming* and *spammers* in Twitter. These specific Twitter users try to increase their number of followers by any means. To achieve this goal, they exploit two relatively straightforward principles based on the reciprocation of the *follow* link:

- **FMIFY** (Follow Me and I Follow You): the user ensures his potential followers that he will follow them back if they follow him first;
- **IFYFM** (I Follow You, Follow Me): on the contrary, the user systematically follows other users, hoping to be followed back.

In their work, Ghosh *et al.* [6] noticed that users responding the most to the solicitations of spammers are real (*i.e.* neither bots nor fake accounts), *active* and even sometimes *popular* users, that they called *social capitalists*. Using this observation, they constituted a list of $100,000$ social capitalists -namely the most responsive ones to the solicitations of spammers. Social capitalists are not healthy for a social networking service, since their methods to gain visibility and influence are not based on the production of relevant content and on getting a higher credibility. From this point of view, their high number of followers can be considered as undeserved, and biases all services based on the assumption that visible users produce or fetch interesting content (e.g. search or recommendation engines).

Using two purely topological measures (and therefore without considering any content), Dugué and Perez [7] designed a method to detect and classify efficiently these users. These measures are based on neighborhood comparisons, namely between the sets of followers $N^-(u)$ (incoming neighbors) and friends $N^+(u)$ (outgoing neighbors) of a user of interest u. The first is called the *overlap index* [8], and is used to detect social capitalists:

$$O(u) = \frac{|N^-(u) \cap N^+(u)|}{\min\{|N^-(u)|, |N^+(u)|\}} \quad (1)$$

Its value ranges from 0 (regular user) to 1 (social capitalist). The second is the *ratio* r, and is used to distinguish between social capitalists using the **FMIFY** ($r \leq 1$) and **IFYFM** ($r > 1$) techniques:

$$r(u) = \frac{|N^+(u)|}{|N^-(u)|} \quad (2)$$

Dugué and Perez [7] also use a third criterion, the number of followers, which corresponds to the incoming degree of the considered node, noted $d^{in}(u)$. They define *low in-degree social capitalists* as social capitalists having less between 500 $10,000$ followers, and *high in-degree social capitalists* as the remaining set of social capitalists. The latter users are considered as successful social capitalists, while the former ones are more popular. It is interesting to notice that in the network we consider, most users with more than $10,000$ followers are social capitalists (70%). Moreover, users with such a number of followers constitute less than 0.1% of the network, which justifies their popularity.

In the experimental part of this article, we decide to use this method to identify the social capitalists in the studied data, instead of the list manually curated by [6]. The reason for this is that the latter seems less exhaustive since it excludes users who do not follow spammers, and does not contain spammers nor bots. Furthermore, some of them have only a few followers, or only a few reciprocate followers-friends links. Finally, the method proposed by Dugué and Perez [7] achieved a greater than 80% accuracy when comparing the social capitalists it detected with those from the list.

III. Identifying Community Roles

In order to characterize the roles of nodes in communities, Guimerà and Amaral [11] defined two complementary measures which allow them to place each node on a 2D role space. Then, they proposed various thresholds to discretize this space, each resulting subspace corresponding to a specific role. We first present this method, then highlight its limitations, and finally propose some solutions to these problems.

A. Original approach

Measures. The two measures are related to the *internal* and *external connectivity* of the node with respect to its community. In other words, they respectively deal with how a node is connected with other nodes inside and outside of its own community. The first measure, called *within-module degree*, is based on the notion of *z-score*. Since the z-score will be used again afterwards, we define it in a generic manner. Let $f(u)$ be any function defined on the nodes, that is f associates a numerical value to any node u of the considered graph. The z-score $Z_f(u)$ w.r.t. the community of u is defined by:

$$Z_f(u) = \frac{f(u) - \mu_i(f)}{\sigma_i(f)}, u \in C_i \quad (3)$$

where C_i stands for a community, and $\mu_i(f)$ and $\sigma_i(f)$ respectively denote the mean and the standard deviation of f over the nodes belonging to community C_i.

Now, let $d_{int}(u)$ be the *internal degree* of a node u, i.e. the number of links u has with nodes belonging to its own community. Then, the *within-module degree* of a node u, denoted $z(u)$ by Guimerà and Amaral [11], corresponds to the z-score of its internal degree. Note that z evaluates the connectivity of a node towards its community with respect to that of the other nodes of the same community.

The second measure, called *participation coefficient*, is defined as follows:

$$P(u) = 1 - \sum_i \left(\frac{d_i(u)}{d(u)}\right)^2 \quad (4)$$

where $d(u)$ denotes the *degree* of the node (*i.e.* the number of links it has towards other nodes), and $d_i(u)$ the *community degree* of u (i.e. the number of links it has towards nodes of community C_i). Note that when C_i corresponds to the community of u, then $d_i(u) = d_{int}(u)$. Roughly speaking, the participation coefficient evaluates the connectivity of a node to the communities. If it is close to 0, then the node is connected to one community only (likely its own). On the contrary, if it is close to 1, then the node is uniformly linked to a large number of communities.

Community Roles. Both measures are used to characterize the *role* of a node within its community. Guimerà and Amaral [11] defined 7 different roles by discretizing the 2D space formed by z and P. They first used a threshold on the within-module degree, which allowed them to distinguish *hubs* (that is, nodes with $z \geqslant 2.5$) from other nodes (called non-hubs). Such hubs are considered as highly linked to their community, when compared to other nodes of the same community. Those two categories are subdivided thanks to several thresholds defined on the participation coefficient (by order of increasing P), as shown in Table I.

Community role				External Connectivity
Within-Module Degree		Participation Coefficient		
Hub	$z \geq 2.5$	Provincial	$P \leq 0.30$	Low
		Connector	$P \in]0.30; 0.75]$	Strong
		Kinless	$P > 0.75$	Very strong
Non-Hub	$z < 2.5$	Ultra-peripheral	$P \leq 0.05$	Very low
		Peripheral	$P \in]0.05; 0.62]$	Low
		Connector	$P \in]0.62; 0.80]$	Strong
		Kinless	$P > 0.80$	Very strong

TABLE I. Classification of roles according to their community-related connectivity.

Directed Variants. Many networks representing real-world systems, such as the Twitter network we study here, are directed. Of course, it is possible to analyze them through the undirected method, but this would result in a loss of information.

Yet, extending these measures is quite straightforward: the standard way of proceeding consists in distinguishing incoming and outgoing links. In our case, this results in using 4 measures instead of 2: in- and out- versions of both the within-module degree and participation coefficient. Let us note d^{in} the in-degree of a node, i.e. the number of incoming links connected to the node. Then one can consider the *internal in-degree* of a node, noted d^{in}_{int}, corresponding to the number of incoming links the node has inside its community. By processing the z-score of this value, one can derive the *within-module in-degree*, noted z^{in}. Let us note d^{in}_i the *community in-degree*, i.e. the number of incoming links a node has from nodes in community C_i. We can now define the *incoming participation coefficient*, noted P^{in}, by substituting d^{in} to d and d^{in}_i to d_i in Equation (4). We similarly define z^{out} and P^{out}, using the outgoing counterparts d^{out}, d^{out}_{int} and d^{out}_i. In the rest of the article, we call this set of measures the *directed variants*, by opposition to the *original measures* of Guimerà and Amaral [11].

B. Limitations of this approach

As mentioned before, we identify two limitations in the approach of Guimerà and Amaral [11]. The first concerns the way the participation coefficient represents the nodes external connectivity, whereas the second is related to the thresholds used for the within-module degree.

External Connectivity. We claim that the external connectivity of a given node, i.e. the way it is connected to communities other than its own, can be precisely described in three ways: first, by considering its *diversity*, i.e. the number of concerned communities ; second, in terms of *intensity*, i.e. the number of external links ; third relatively to its *heterogeneity*, i.e. the distribution of external links over communities. The participation

coefficient combines several of these aspects, mainly focusing on heterogeneity, which lowers its discriminant power. This is illustrated in Figure 1: the external connectivity of the central node is very different in each one of the presented situations. However, P is the same in all cases.

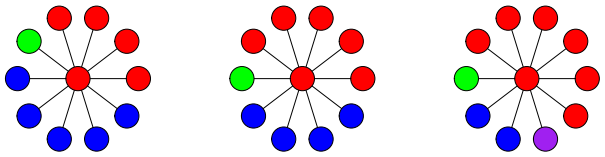

Fig. 1. Each pattern represents a community. In each case, the participation of the central node is 0.58.

In order to be more illustrative, let us consider two users from our data, which have the same community role according to the original measures. We select two nodes both having a z greater than 2.5 and a P close to 0.25. So according to Guimerà and Amaral [11] (see Table I), they both are provincial hubs, and should have a similar behavior w.r.t. the community structure of the network. However, let us now point out that the first user is connected to 50 nodes outside its community, whereas the second one has $200,000$ connections. This means they actually play different roles in the community structure, either because the second one is connected to much more communities than the first one, or because its number of links towards external communities is much larger than for the first user. Similar observations can be made for the directed variants of the participation coefficient. The measures used to define the external connectivity should take this difference into account and assign different roles to these nodes.

Fixed Thresholds. As indicated in the supplementary discussion of Guimerà and Amaral [11], the thresholds originally used to identify the roles were obtained empirically. They first processed P and z for different types of data: metabolic, proteome, transportation, collaboration, computer and random networks. Then, they detected basins of attraction, corresponding to regularities observed over all the studied networks. Each role mentioned earlier corresponds to one of these basins, and the thresholds were obtained by estimating their boundaries.

Implicitly, these thresholds are supposed to be universal, but this can be criticized. First, Guimerà and Amaral [11] used only one community detection method. A different community detection method can lead to a different community structure, and therefore possibly different basins of attraction. Furthermore, z is not normalized, in the sense it has no fixed boundaries. There is no guarantee the threshold originally defined for this measure will stay meaningful on other networks. The values obtained for z in our experiments are far higher for some nodes than the ones observed by Guimerà and Amaral [11]. We also observe that the proportion of nodes considered as hubs (i.e. $z \geq 2.5$) by Guimerà and Amaral [11] is much smaller in our network than in the networks they consider: 0.35% in ours versus 2% in theirs. These thresholds seem to be at least sensitive either to the size of the data, the structure of the network, or to the community detection method.

It is therefore necessary to process new thresholds, more appropriate to the considered data. However, the method used by Guimerà and Amaral [11] itself is difficult to apply, because it requires a lot of data. We now present how to overcome these limitations.

C. Proposed Approach

Generalized Measures. In place of the single participation coefficient, we propose 3 new measures aiming at representing separately the aspects of external connectivity: *diversity*, *intensity* and *heterogeneity*. Moreover, a fourth measure is used to describe the internal connectivity.

Because we deal with directed links, each one of these measures exists in two versions: incoming and outgoing (as explained in section III-A), resulting in 8 effective measures. However, for simplicity matters, we ignore link directions when presenting them in the rest of this section.

All our measures are expressed as z-scores. We know community sizes are generally power-law distributed, as described in [14], which means their sizes are heterogeneous. Our community-based z-scores (cf. Equation (3)) allows to normalize the measures relatively to the community size, and therefore to take this heterogeneity into account.

Diversity. The *diversity* $D(u)$ evaluates the number of communities to which a node u is connected (other than its own), w.r.t. the other nodes of its community. This measure does not take into account the number of links u has to each community. Let $\epsilon(u)$ be the number of external communities to which u is connected. The diversity is defined as the z-score of ϵ w.r.t. the community of u. It is thus obtained by substituting ϵ to f in Equation (3).

External intensity. The *external intensity* $I_{ext}(u)$ of a node u measures the amount of links u has towards communities other than its own, w.r.t. the other nodes of its community. Let $d_{ext}(u)$ be the external degree of u, that is the number of links u has with nodes belonging to another community than its own. The external intensity is defined as the z-score of the external degree, i.e. we obtain it by substituting d_{ext} to f in Equation (3).

Heterogeneity. The *heterogeneity* $H(u)$ of a node u measures the variation of the number of links a node u has, from one community to another. To that aim, we compute the standard deviation of the number of links u has to each community. We denote this value by $\delta(u)$. The heterogeneity is thus the z-score of δ w.r.t. the community of u. As previously, it can be obtained by substituting δ to f in Equation (3).

Internal intensity. In order to represent the internal connectivity of the node u, we use the z measure of Guimerà and Amaral [11]. Indeed, it is based on the notion of z-score, and is thus consistent with our other measures. Moreover, we do not need to add measures such as diversity or heterogeneity, since we consider one node can belong only to one community. Due to the symmetry of this measure with the external intensity, we refer to z as the *internal intensity*, and denote it by $I_{int}(u)$.

Unsupervised Role Identification. Our second modification concerns the way roles are defined. As mentioned before, the thresholds defined by Guimerà and Amaral [11] are not necessarily valid for all data. Moreover, our generalization of the measures invalidates the existing thresholds, since we

have now 8 distinct measures, all different from the original ones. We could try estimating more appropriate thresholds, but as explained in section III-B, the method originally used by Guimerà and Amaral [11] to estimate their thresholds is impractical. The fact our measures are all z-scores also weakens the possibility to get thresholds applicable to all systems, which means the estimation process should potentially be performed again for each studied system.

To overcome these problems, we propose to apply an automatic method instead, by using unsupervised classification. First, we process all the measures for the considered data. Then, a cluster analysis method is applied. Each one of the clusters identified in the measure space is considered as a community role. This method is not affected by the number of measures used, and amounts to adjusting thresholds to the studied system. If the number of roles is known in advance, for instance because of some properties of the studied system, then one can use an appropriate clustering method such as k-means, which allows specifying the number k of clusters to find. Otherwise, it is possible to use cluster quality measures to determine which k is the most appropriate ; or to apply directly a method able to estimate at the same time the optimal number of clusters and the clusters themselves.

IV. Community Roles of Social Capitalists

A. Data and Tools

We analyze a freely-available anonymized Twitter network, collected in 2009 by Cha *et al*. [13]. It contains about 55 million nodes representing Twitter users, and almost 2 billion directed links corresponding to friend-to-follower relationships. We had to consider the size of these data when selecting our analysis tools. For community detection, we selected the Louvain method [15], because it is widespread and proved to be very efficient when dealing with large networks. We retrieved the C++ source code published by its authors, and adapted it in order to optimize the directed version of the modularity measure, as defined by Leicht and Newman [16]. All the role measures, that is Guimerà and Amaral's original measures, their directed variants (section III-A) and our new measures (section III-C), were computed using the community structure detected through this means. We also implemented them in C++, using the same sparse matrix data structure than the one used in the Louvain method.

All resulting values were normalized, in order to avoid scale difference problems when conducting the cluster analysis. Since we do not know the expected number of roles, the clustering was performed using an open source implementation of a distributed version of k-means [17]. Indeed, centralized versions are based on a unique distance matrix, and turned out to be too demanding in terms of memory. We applied this algorithm for k ranging from 2 to 15, and selected the best partition in terms of Davies-Bouldin index [18]. We selected this index because it is a good compromise between the reliability of the estimated quality of the clusters, and the computing time it requires. All pre- and post-processing scripts related to the cluster analysis were implemented in R. The whole source code is available at the following address: https://github.com/CompNet/Orleans

B. Roles Expected for Social Capitalists

We expect the degree of social capitalists to play an important role considering their position (see Section II). High in-degree social capitalists (namely greater than $10,000$) should be well connected to their communities -hubs- or to the other communities -connectors, or both. Being connectors would indicate they obtained a high visibility on the whole network and not only in their own communities. Furthermore, because we take the direction of links into account in our measures, we expect social capitalists to be discriminated according to their ratio, i.e. the number of outgoing links divided by the number of incoming links. We especially expect high in-degree social capitalists with a small ratio (so-called *passive social capitalists* according to [7]) to be highly connected to their communities and to the rest of the graph. Considering low degree social capitalists, it is not possible to predict their roles without any further information. The study will thus be of great interest to characterize their visibility.

C. Detected Roles

For the sake of completeness, we first used the original undirected measures of Guimerà and Amaral [11]. We obtained only 2 roles, each one concerning too many nodes to bring up any valuable information regarding the studied system. Since this might be due to the fact these measures ignore link directions, we then worked with their directed variants (section III-A), and then with our generalized measures (section III-C). In both cases we used the unsupervised role identification method we proposed (section III-C).

Directed Variants. A correlation study shows z^{out} and z^{in} are slightly correlated (with a correlation coefficient $\rho < 0.3$), whereas the correlation is zero for all other pairs of measures. This seems to confirm the interest of considering link directions in the role measures. When doing the cluster analysis, the most separated clusters are obtained for $k = 6$. An ANOVA followed by *post hoc* tests (t-test with Bonferroni's correction) showed significant differences exist between all clusters and for all measures.

An analysis of the distribution of high in-degree social capitalists in these clusters shows that a few of these users occupy a connector hub role. This is quite expected as said in IV-B. However, most of the high degree social capitalists are considered as non-hubs and peripheral or ultra-peripheral nodes. More than 60% of the users with a high ratio are classified as ultra-peripheral nodes for both incoming and outgoing directions, which is rather surprising since they have a really high degree. However, they are classified in a cluster with low z and P (both in- and out- versions). The low z indicates these users are not much connected to their community (relatively to the other nodes of the same community), and must thus be more connected to other communities. Still, P does not highlight this aspect of their community-related connectivity, and they appear as peripheral. This inconsistency of the detected roles confirms the limitations of P described in section III-B.

Generalized Measures. Most generalized measures are slightly correlated, with values ranging from almost 0 to 0.4. In particular, both versions of the same measure (incoming vs. outgoing) are only slightly correlated, which is another

confirmation of the interest of considering link directions. Only three measures are strongly correlated: internal and external intensities and heterogeneity (ρ ranging from 0.78 to 0.92). The relation between both intensities seems to indicate that variations on the total degree globally affect similarly internal and external degrees. The very strong correlation observed between heterogeneity and intensity means only nodes with low intensity are homogeneously connected to external communities, whereas nodes with many links are connected heterogeneously.

Similarly to the directed measures, the most separated clusters are obtained with $k = 6$. These 6 clusters are given in Table II with their sizes and roles. However, the correspondance with the original nomenclature is rougher, since these measures are farther from the original ones. The average of each measure per cluster is showed in Table III. Like before, ANOVA and *post hoc* tests showed significant differences between all clusters and for all measures. We now conduct a detailed analysis of the different roles we obtain.

Cluster	Size	Proportion	Role
1	24543667	46.68%	Ultra-peripheral non-hubs
2	304	$< 0.01\%$	Kinless hubs
3	303674	0.58%	Connector hubs
4	11929722	22.69%	Incoming Peripheral non-hubs
5	10828599	20.59%	Outgoing Peripheral non-hubs
6	4973717	9.46%	Connector non-hubs

TABLE II. CLUSTERS DETECTED WITH THE GENERALIZED MEASURES: SIZES IN TERMS OF NODE COUNT AND PROPORTION OF THE WHOLE NETWORK, AND ROLES ACCORDING TO THE GUIMERÀ AND AMARAL [11] NOMENCLATURE.

Cluster 1. Because both internal intensity versions (equivalent to z) are negative, nodes in this cluster cannot be hubs. The negative external measures indicate these nodes are not connectors either. We can thus consider them as ultra-peripheral non-hubs. This cluster is the largest one, with 47% of the network nodes. This confirms the matching with this role, whose nodes constitute generally most of the network.

Clusters 4 and 5. Cluster 4 is very similar to Cluster 1. However, its incoming diversity is 0.69. These nodes are again peripheral, because the external intensity is negative. Still, incoming links come from a larger number of communities. Cluster 5 is also similar to Cluster 1. However, both versions of diversity are positive for this cluster, with an outgoing diversity of 0.60. External links are thus connected to a larger number of communities. Clusters 4 and 5 are the second (23%) and third (21%) largest ones, respectively. By gathering all the peripheral and ultra-peripheral nodes, we obtain 91% nodes of the network.

Cluster 6. The internal intensity is still close to 0 but positive. Thus, these nodes are non-hubs, even if they are more connected to their community than those of the previous clusters. Like the other external measures, the external intensity is low but still positive. These nodes are relatively well-connected to other communities, and we can therefore consider them as connectors. Both versions of the diversity are relatively high, which indicates these nodes are not only more connected to their community as well as others, but also to a larger number of distinct communities.

Cluster 3. The high internal intensity allows us to state that these nodes are hubs. Furthermore, the high external measures indicate these nodes are connected to a high number of nodes from a lot of other communities, and thus are connector hubs. Notice outgoing measures are higher. This cluster represents only 0.6% of the network, meaning this role is very uncommon.

Cluster 2. This observation is even more valid for Cluster 2, which represents much less than 1% of the nodes. For this cluster, all measures are really high. The incoming versions are always higher than their outgoing counterparts. We call these users kinless hubs according to Guimerà and Amaral's nomenclature.

Cluster	$\mathbf{I}^{\mathbf{out}}_{\mathbf{int}}$	$\mathbf{I}^{\mathbf{in}}_{\mathbf{int}}$	$\mathbf{D}^{\mathbf{out}}$	$\mathbf{D}^{\mathbf{in}}$
1	-0.12	-0.03	-0.55	-0.80
2	94.22	311.27	7.18	88.40
3	5.52	1.40	5.60	3.10
4	-0.04	0.00	-0.37	0.69
5	-0.03	-0.01	0.60	0.19
6	0.48	0.12	1.96	1.70

Cluster	$\mathbf{I}^{\mathbf{out}}_{\mathbf{ext}}$	$\mathbf{I}^{\mathbf{in}}_{\mathbf{ext}}$	$\mathbf{H}^{\mathbf{out}}$	$\mathbf{H}^{\mathbf{in}}$
1	-0.09	-0.04	-0.12	-0.06
2	113.87	283.79	112.79	285.57
3	5.28	1.43	6.76	2.34
4	-0.07	0.00	-0.10	-0.01
5	-0.03	-0.02	-0.04	-0.02
6	0.35	0.12	0.53	0.19

TABLE III. AVERAGE GENERALIZED MEASURES OBTAINED FOR THE 6 DETECTED CLUSTERS .

It is worth noticing that, whatever the considered measures, some of the roles defined by Guimerà and Amaral [11] are not represented in the studied network. This is consistent with the remarks previously made for other data by Guimerà and Amaral [11], and confirms the necessity of having an unsupervised approach to define roles in function of measures. It is also consistent with the strong correlation observed between internal and external intensities: missing roles would be nodes possessing a high internal intensity but a low external one, or vice-versa. However, those are very infrequent in our network.

D. Relations between clusters

We now discuss how the nodes are connected depending on the role they hold. Figure 2 is a simplified representation of this interconnection pattern.

The outgoing links of ultra-peripheral (Cluster 1) and peripheral (Clusters 4 and 5) nodes target mainly kinless hubs (Cluster 2) and connectors (Clusters 3 and 6), representing 74% (Cluster 1), 82% (Cluster 4), and 74% (Cluster 5) of their connections. These (ultra-)peripheral nodes, which are the most frequent in the network, thus mainly follow very connected users, probably the most influent and relevant ones. This seems consistant: they follow only a few users, and so choose the most visible ones.

Connector nodes (Clusters 3 and 6) are mainly linked to other connectors nodes. They have the tightest connection, since their arcs amounts to a total of 43% of the network links. This is worth noticing, because these clusters are far from being the largest ones. They are also largely connected to the rest of the clusters too, especially with outgoing links.

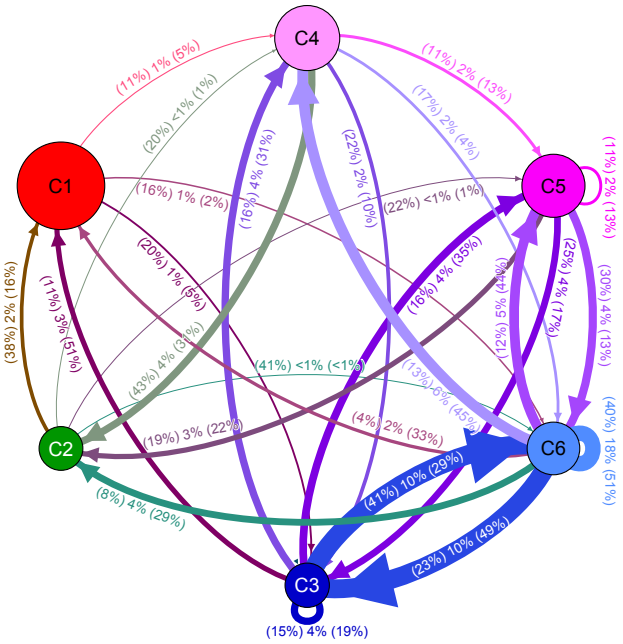

Fig. 2. Interconnection between clusters. A vertex i corresponds to Cluster i from Table II. An arc (i, j) represents the set of links connecting nodes from Cluster i to nodes from Cluster j. It is labeled with 3 values, each one describing which proportion of links the arc represents, relatively to 3 distinct sets: first relatively to all links starting from Cluster i, second relatively to all links in the whole network, and third relatively to all links ending in Cluster j. The arc thickness is proportional to the second value. For matters of readability, arcs representing less than 1% of the network links are not displayed.

Connectors follow massively users of all clusters, so we suppose they constitute the backbone of the network.

Kinless hubs (Cluster 2) are massively followed by non-hubs, representing 38% (Cluster 1), 43% (Cluster 4), 19% (Cluster 5) and 8% (Cluster 6) of their outgoing links. And interestingly, the links coming from kinless hubs target the same clusters: 9% go to Cluster 1, 20% to Cluster 4, 22% to Cluster 5 and 41% to Cluster 6. This means the most visible and popular nodes of the network mostly follow and are followed by much less popular users. One could have expected the network to be hierarchically organized around roles, with more peripheral nodes connected to less peripheral nodes. But this is clearly not the case. First, (ultra-)peripheral nodes are marginally connected to other nodes holding the same role, they prefer to follow connectors and/or hubs. Second, kinless and connector hubs, although well connected to connector non-hubs, do not have direct links, i.e. these users do not follow each other.

E. Position of Social Capitalists

As stated previously, we use a list of approximately $160,000$ social capitalists as detected by Dugué and Perez [7]. In the following, we analyze how social capitalists are distributed amongst the detected roles. As explained section II, we split social capitalists according to their in-degree (number of followers). Recall that *low in-degree social capitalists* have an in-degree between 500 and $10,000$, and *high in-degree social capitalists* an in-degree greater than $10,000$. These social capitalists are known for having especially well succeeded in their goal of gaining visibility.

The tables in this section describe how the various types of social capitalists are distributed over the clusters. In each cell, the first row is the proportion of social capitalists belonging to the corresponding cluster, and the second one is the proportion of cluster nodes which are social capitalists. Values of interest are indicated in bold and discussed in the text.

Low in-degree social capitalists. Low in-degree social capitalists are mostly assigned to three clusters: 3, 5 and 6 (see Table IV). Most of them belong to Cluster 6, which contains non-hub connector nodes. These nodes, which have only slightly more external connections than the others, are nevertheless connected to far more communities. Social capitalists in this cluster seem to have applied a specific strategy consisting in creating links with many communities. This strategy is still not completely working, though, as shown by the relatively low external incoming intensity (meaning they do not have that many followers).

Nodes from Cluster 3 are connector hubs, who follow more users than the others. Because **IFYFM** social capitalists have a ratio greater than 1 and thus more friends than followers, it is quite intuitive to observe that they are twice as many than the other users in this cluster. The high outgoing diversity of Cluster 3 tells us that these social capitalists follow users from a large variety of communities, not only theirs (to which they are well connected). The high external outgoing intensity show that these users massively engage in the **IFYFM** process, but did not yet receive a lot of following back, as shown by their low external incoming intensity. Finally, roughly 20% of social capitalists with ratio $r < 1$ belong to Cluster 5, which contains non-hub peripheral nodes. This shows that a non-neglictible share of social capitalists are isolated relatively to both their community and the other ones.

Ratio	Cluster 1	Cluster 2	Cluster 3
$r \leq 1$	0.01% $< 0.01\%$	0.00% 0.00%	$\mathbf{23.10\%}$ 3.71%
$r > 1$	0.03% $< 0.01\%$	0.00% 0.00%	$\mathbf{18.78\%}$ $\mathbf{6.61\%}$

Ratio	Cluster 4	Cluster 5	Cluster 6
$r \leq 1$	3.42% 0.14%	$\mathbf{18.28\%}$ 0.08%	$\mathbf{55.19\%}$ 0.54%
$r > 1$	0.48% $< 0.01\%$	$\mathbf{14.31\%}$ 0.14%	$\mathbf{66.40\%}$ 1.43%

TABLE IV. DISTRIBUTION OF LOW IN-DEGREE SOCIAL CAPITALISTS OVER CLUSTERS OBTAINED FROM THE GENERALIZED MEASURES.

These observations show that most of these users are deeply engaged in a process of soliciting users from other communities, not only theirs. Some of them are even massively following users from a wide diversity of communities. This tends to show that these users may obtain an actual visibility across many communities of the network by spreading their links efficiently.

High in-degree social capitalists. Most of the high in-degree social capitalists are gathered in Cluster 3 (see Table V), corresponding to connector hubs. This is consistent with the fact these users have a high degree. Users of Cluster 3 have a high outgoing diversity and a high outgoing external intensity: this shows they practice the **IFYFM** strategy actively, by

following a lot of users from a wide range of communities. The rest of these users is contained in Cluster 2. Nodes in these clusters are kinless hubs and thus can be considered as successful users. Indeed, they are massively followed by a very high number of users from an extremely large variety of communities. Only high-degree social capitalists with a ratio smaller than 0.7 and a few with a ratio smaller than 1 are classified in this cluster. This is consistent with the roles one could expect for social capitalists (section II).

Ratio	Cluster 1	Cluster 2	Cluster 3
$r \leq 0.7$	0.00%	**12.14%**	**87.29%**
	0.00%	**21.05%**	0.15%
$0.7 < r \leq 1$	0.00%	1.55%	**95.64%**
	0.00%	**7.24%**	0.45%
$r > 1$	0.00%	0.03%	**97.99%**
	0.00%	0.33%	1.22%

Ratio	Cluster 4	Cluster 5	Cluster 6
$r \leq 0.7$	0.00%	0.00%	0.57%
	0.00%	0.00%	$< 0.01\%$
$0.7 < r \leq 1$	0.00%	0.00%	2.81%
	0.00%	0.00%	$< 0.01\%$
$r > 1$	0.00%	0.00%	1.98
	0.00%	0.00%	$< 0.01\%$

TABLE V. DISTRIBUTION OF HIGH IN-DEGREE SOCIAL CAPITALISTS OVER CLUSTERS OBTAINED FROM THE GENERALIZED MEASURES.

These observations mean that most of these users are well connected in their communities but also with the rest of the network. This shows the efficiency of these users strategies. Indeed, most of the users are linked to a wide range of communities, and thus reach a high visibility in a large part of the network.

V. CONCLUSION

In this article, our goal is to characterize the position of social capitalists in Twitter. For this purpose, we propose an extension of the method defined by Guimerà and Amaral [11] to characterize the community role of nodes in complex networks. We first define directed variants of the original measures, and extend them further in order to take into account the different aspects of node connectivity. Then, we propose an unsupervised method to determine roles based on these measures. It has the advantage of being independant from the studied system. Finally, we apply our tools to a friend-to-follower Twitter network. We find out the different kinds of social capitalists occupy very specific roles. Those of low in-degree are mostly connectors non-hubs. This shows they are engaged in a process of spreading links across the whole network, and not only their own community. Those of high in-degree are classified as kinless or connectors hubs, depending on their ratio r. This shows the efficiency of their strategies, which lead to a high visibility for a vast part of the network, not only for their own community.

The most direct perspective for our work is to assess its robustness. In particular, it is important to know how the stability of the detected communities and clusters affects the identified roles. In this study, the very large size of the data prevented us to do so: first, it was a strong constraint when selecting the tools we used for community detection and cluster analysis, and second it was not possible to repeat these processing many times to evaluate the stability of their results. We plan to work on this point by using smaller datasets. On a related note, we want to apply our method to other systems, in order to check for its general relevance. The method itself can also be extended in two ways. First, it would be relatively straightforward to take link weights into account (although this was not needed for this work). Second, and more interestingly, it is also possible to adapt it to overlapping communities (by opposition to the mutually exclusive communities considered in this work) in a very natural way, by introducing additional internal measures symmetrical to the existing external ones. This could be a very useful modification when studying social networks, since those are supposed to possess this kind of community structures, in which a node can belong to several communities at once [19].

REFERENCES

[1] B. Bosker. (2011) Twitter: We now have over 200 million accounts. Huffington Post. [Online]. Available: http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html

[2] R. Holt. (2013) Twitter in numbers. The Telegraph. [Online]. Available: http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html

[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *WebKDD/SNAKDD*, 2007, pp. 56–65.

[4] S. Rodgers. (2013, August) Behind the numbers: how to understand big moments on Twitter. Twitter. [Online]. Available: https://blog.twitter.com/2013/behind-the-numbers-how-to-understand-big-moments-on-twitter

[5] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network," in *SCA'10*, 2010, pp. 177–184.

[6] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. Gummadi, "Understanding and combating link farming in the twitter social network," in *WWW*, 2012, pp. 61–70.

[7] N. Dugué and A. Perez, "Social capitalists on Twitter: detection, evolution and behavioral analysis," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1–15, 2014.

[8] G. Gaylord Simpson, "Mammals and the nature of continents," *Am. J. of Science*, no. 241, pp. 1–41, 1943.

[9] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, 2004.

[10] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75–174, Feb 2010.

[11] R. Guimerà and L. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, pp. 895–900, 2005.

[12] J. Scripps, P.-N. Tan, and A.-H. Esfahanian, "Node roles and community structure in networks," in *WebKDD/SNAKDD*, 2007, pp. 26–35.

[13] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *ICWSM*, 2010.

[14] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, "Characterizing the community structure of complex networks," *PLoS ONE*, vol. 5, no. 8, p. e11976, 2010.

[15] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 10, p. P10008, 2008.

[16] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Phys. Rev. Lett.*, vol. 100, no. 11, p. 118703, 2008.

[17] W.-K. Liao. (2009, Oct) Parallel k-means data clustering. Northwestern University. [Online]. Available: http://users.eecs.northwestern.edu/~wkliao/Kmeans/index.html

[18] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979.

[19] S. Arora, R. Ge, S. Sachdeva, and G. Schoenebeck, "Finding overlapping communities in social networks: Toward a rigorous approach," in *EC'12*, 2012.