González-Bailón et al. *EPJ Data Science* 2014, **3**:32
http://www.epjdatascience.com/content/3/1/32

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

The emergence of roles in large-scale networks of communication

Sandra González-Bailón[1*], Ning Wang[2] and Javier Borge-Holthoefer[3*]

*Correspondence:
sgonzalezbailon@asc.upenn.edu;
jborge@qf.org.qa
[1]Annenberg School for Communication, University of Pennsylvania, Philadelphia, USA
[3]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
Full list of author information is available at the end of the article

Abstract

Communication through social media mediates coordination and information diffusion across a range of social settings. However, online networks are large and complex, and their analysis requires new methods to summarize their structure and identify nodes holding relevant positions. We propose a method that generalizes the sociological theory of brokerage, originally devised on the basis of local transitivity and paths of length two, to make it applicable to larger, more complex structures. Our method makes use of the modular structure of networks to define brokerage at the local and global levels. We test the method with two different data sets. The findings show that our approach is better at capturing role differences than alternative approaches that only consider local or global network features.

Keywords: modularity; bridges; structural holes; structural similarity; online networks

1 Introduction

Networks of inter-personal communication have grown larger and more complex with the emergence of digital technologies [1]. Methods that summarize this intricacy help identify the building blocks that explain network dynamics like diffusion. Much research has been conducted in the last two decades to assess how the structure of networks correlates with their dynamics. Several methods have been developed to reduce the complexity of networks through the identification of their characteristic features, which involves determining the differences between observed and expected patterns under a null model of random connections. Reduction techniques like community detection or backbone extraction belong to this tradition; both techniques use the distribution of edges as the basis from where to identify the inner structure of a network [2, 3].

Blockmodels and structural equivalence offer an alternative approach, shifting attention from the edges to the nodes and to how similar they are as assessed by their connections to other nodes in the network [4–6]. The notion of structural equivalence has a long history in the analysis of social networks and the definition of roles, but it has recently been extended to make it less strict in its technical definition, more scalable, and applicable to other complex (non-social) networks [7]. Devising methods for role identification is important for network science because roles offer a scheme for network reduction and the construction of simplified maps of the original structure; but also, and significantly, because they offer a criterion to group nodes in categories that might be associated with

Springer

© 2014 González-Bailón et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

similar behavior. We follow this prior research and propose a new method for the identification of roles in large networks.

The main assumption of our work is that roles respond to a division of labor that reflects different functions, or behavior, within the network. Detecting structurally similar positions depends on the network features emphasized by the method. Our goal is to compare different methods and assess which structural features are more successful at identifying relevant nodes. We use two different empirical data sets for that endeavor: the first, collected from Twitter, tracks patterns of communication around a political protest organized in May of 2012. The second, smaller and collected manually, is the classic (and renowned) Zachary's Karate club network [8], which we use mostly as a robustness test for our method. We analyze these data with two aims: to determine whether users that shared similar network positions (according to alternative methods) behaved similarly in the exchange of information; and to identify the roles that were more significant in that exchange.

On a theoretical level, we want to add a mesoscopic dimension to classic theories of brokerage in social networks, which have traditionally focused on local information flows [9–12]. As the proponents of this previous work, we sustain that roles are theoretically important because they are independent of the specific network under analysis: roles help us draw a typology of actors that transcends the composition of a given network. Unlike that previous research, however, we also contend that large networks require identifying roles on a coarse-grained level of analysis that takes the entire network structure into account; the relevant source of actor heterogeneity is otherwise lost in the detail of local positions. This is particularly important in the study of collective phenomena like protest coordination via social media: these communication networks tend to be large and complex, with distinct local affiliations, a small core of highly active actors, and a large periphery of lowly committed users [13–16]. The ability to broker information in such networks does not arise from local connectivity alone.

The rest of the paper proceeds as follows. First, we revise previous work on roles in social networks and defend the need to follow a more mesoscopic approach to role identification. Then we discuss our data and methods, emphasizing that we can generalize our analytical strategy to other data sets and case studies. The presentation of findings considers whether actors that are classified as structurally similar exhibit also similar behavior. We compare three approaches to the identification of roles, and we evaluate their performance when it comes to capturing differences in communication activity. The findings show that our method, which defines structural similarity in terms of local and global brokerage, is better able to capture differences in communication dynamics. A summary and discussion of the main findings closes the paper.

2 Roles in social networks

Sociologists have long characterized individual actors by the roles they play in social systems [17]. Roles refer to a set of expectations and patterns of behavior associated to positions in a social structure. The network translation of this idea usually relies on the notion of structural equivalence and its looser version, structural similarity: actors playing similar roles will have similar patterns of connections with other actors in a network [6, 18–20]. Two structurally similar actors have comparable ties to each other and to the other actors in the network. Roles offer an abstraction of these similarities: they provide categories that reduce networks to comparable building blocks.

Traditionally, research on social networks has assumed that network structure mirrors preexisting roles: being a professor at a university or a director in a company translates into having a set of similar ties with the other actors in the network (i.e. students and faculty, or other directors). Networks reflect, in this sense, formal roles that are defined within the parameters of shared institutions and mutual expectations. But roles can also emerge spontaneously from interactions in networks: some actors are able to play a differential role because of their structural position, for instance as when brokers emerge in informal networks of communication [10, 11]. According to this idea, ties do not reflect preexisting roles; instead, they allow actors to play certain functions - for instance, broker information. Actors playing a brokering role share a common trait: they span structural holes in a network by building bridges across sub-graphs that would be unconnected in their absence. This position gives them the possibility to control information flows, a role they share with other brokers in the network.

Although the idea of structural similarity requires taking into account connections at the level of the overall network, most network measures of roles (and brokerage in particular) still rely on local patterns. The identification of brokers in social networks is based either on structural constraint [10], which roughly measures transitivity in personal networks; or on the betweenness of actors in paths of length two [11], that is, on being the vertex in an open triad through which the two unconnected nodes interact. In addition to using only local information, these definitions of social roles are purely structural - that is, they capture the potential for brokerage given the absence of alternative routes through which information can flow, but they do not assess the extent to which information actually flows [21, 22]. In other words, roles are defined in terms of positions in a structure, and they are rarely validated in terms of behavior or action.

In spite of these acknowledged limitations, the local approach has still proved fruitful to identify general roles and analyze their association with the performance of organizations and elites [9–12, 23]. Gould and Fernandez [9], for instance, offer a typology of brokers on the basis of how actors mediate communication across pre-defined groups. Their typology (GF from now on) is depicted in Figure 1, panel A. According to this classification, when all nodes in a path of length two (or open triad) are part of the same group, the actor brokering the connection acts as a coordinator; when all nodes belong to different groups, the broker fulfills a liaison role; she acts as a gatekeeper or a representative depending

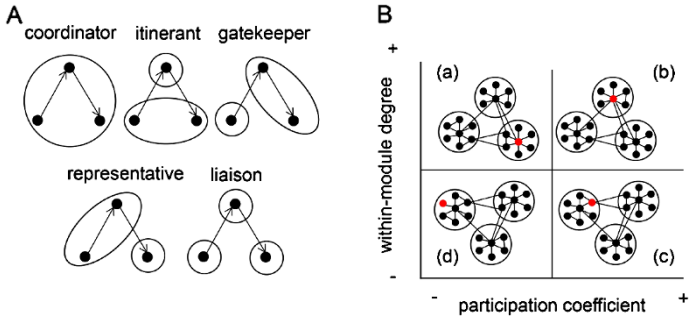


Figure 1 Role schemes derived from connectivity within and across groups. Schematic representation of role classifications derived from two methods. The method illustrated in panel **A** uses a local measure of brokerage [9]. The method illustrated in panel **B** uses the modular structure of a network to define within module degree and the diversity of connections to other modules [6].

on whether she belongs to the same group as the recipient of the information or as the source; and she plays an itinerant role when she belongs to a group distinct from the other two nodes in the triad. Differentiating these roles is important because they highlight the empirical possibility that the effects of brokerage, or its value, change across scenarios - a possibility that is more likely if nodes affiliated to different groups manage different types of information.

In larger networks, using paths of length two as the main criterion to identify structurally similar positions does not offer much space for discrimination - these local configurations abound and most nodes will be part of many of them. This is one of the reasons why a more scalable criterion to identify roles in complex networks was proposed recently by Guimerà and Amaral (GA from now on, [6]). As with the GF typology, the GA method requires a partition of the network in groups, often drawn from the topology itself using community detection methods [2]. Once this partition is generated, nodes are classified in a two-dimensional space that measures their intra-module centrality (i.e. the number of connections with other nodes classified in the same group) and their participation coefficient (i.e. the propensity to connect to nodes that are classified in other groups). This two-dimensional space is then fragmented in regions according to some heuristic derived from how nodes distribute in the space. In general, though, there are four main types of nodes, summarized in Figure 1, panel B: (a) provincial hubs, or nodes that are highly connected with other nodes in their module, but poorly connected to other modules; (b) connector hubs, or nodes that are highly connected to other nodes in their module and are also well connected to other groups; (c) non-hub connectors, or nodes that are poorly connected to their module but well connected to other modules; and (d) peripheral nodes, or nodes that are poorly connected to both their module and other groups.

The two role schemes summarized in Figure 1 overlap to some extent: brokers classified as coordinators are more likely to appear in categories (a) or (d), whereas brokers playing a liaison role are more likely to appear in categories (b) and (c). There are two important differences between the two schemes, however. The first is that the GA approach allows us to make a global definition of what counts as a structural hole based on the relative absence of ties across groups (as identified by community detection methods); this is more discriminatory than the local definition of holes based on paths of length two. In doing so, though, the GA method disregards the importance of local connections: it is personal networks that help us identify the role that actors play in their immediate social environments; the ability to broker connections on that level of analysis can also be consequential in emergent group dynamics [10, 17]. The second difference is that the GA method disregards the directionality of connections, but directionality is crucial in communication networks - and well captured by the GF schema. In the following section, we propose a third hybrid method that integrates the two schemes (HM from here on), and we assess its relative performance for the identification of similarly behaving nodes.

3 Data and methods

We use two data sets to compare the three role detection methods. The first was collected from Twitter for the period April 30 to May 30 2012 using the platform's search API. We applied filtering parameters to select messages that contained hashtags related to the Occupy and 'indignados' movements, retrieving about 445,000 messages. The filtering parameters and observation window were selected with the goal of analyzing communication dynamics around an international call for action, a political protest that took place

in May 12. Using the unique author identifiers contained in the messages, we run further queries to reconstruct the network of following/followers. This one-step snowball crawl returned a network of more than 38 million users; of these, we only retained users that had sent at least one message during the observation period, as well as their connections to other users that were also involved in protest-related communication. In addition, we parsed the messages to identify re-tweets (RTs) and mentions (@), which allowed us to reconstruct direct interactions and explicit channels of information flow amongst users. More information about the data can be found in [24], and a discussion of the sampling procedure in [25].

Sampling choices and defining the boundaries for data collection are always important parameters in any research design. On the basis of prior work assessing the bias in samples of Twitter networks [25] we believe that we are probably underestimating the number of users in the sparser parts of the network (i.e. the periphery) and, as a consequence, also the centrality of the most central users: they get many of their connections from peripheral users (whose individual contributions are small, but statistically large when aggregated). An important aspect of the data for our purposes is that it captures a topology of connections (the follower structure) and a more dynamic layer of actual communication and information exchange (via mentions and RTs). This information is difficult to obtain from other sources of data - including our second data set.

The second network we analyze was collected in the context of anthropological research on conflict dynamics in small groups [8]. This network, which tracks communication among the members of a Karate club, is a well-known benchmark in community detection research because it can be characterized by the existence of two large groups or communities - the two factions that ended up splitting and resulted in the foundation of a new club. This network is small ($N = 34$) and because of that - and the anthropological research providing contextual information - it offers a more intuitive validity test for role detection methods. We use this data to examine the extent to which our method can be generalized across data sources and social settings.

The first step in our analyses involved finding out the modular structure of the two networks (follower structure for the Twitter data). We did so applying the fast greedy community detection method [26], as implemented in the open-source library *igraph* [27]. The method seeks the optimization of modularity Q defined as

$$Q = \frac{1}{2m}\sum_{i,j}[A_{i,j} - P_{i,j}]\delta_{C_iC_j} = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_ik_j}{2m}\right]\delta_{C_iC_j}, \quad (1)$$

where the graph is represented as a matrix with cells $A_{i,j}$, which are valued 1 if a link exists between i and j, 0 otherwise; the second term corresponds to the configuration model, that is, the expected counts under the assumption of random connections between the nodes; finally, Kronecker delta $\delta_{C_iC_j}$ has a value 1 if nodes i and j belong to the same community C, and 0 otherwise.

The fast greedy approach to graph optimal partitioning relies on a recursive agglomerative scheme. As such, all nodes belong to their own community at start time ($M = N$); then the algorithm attempts to merge communities into larger ones. If the resulting partition improves the previous Q value, the merge is accepted and the algorithm advances towards further attempted mergers. Compared to other algorithms to optimize Q (and detect some

sort of modular structure), the greedy approach is efficient and reliable: it can handle very large networks and it typically yields a number of modules $M \ll N$, which fits well the nature of our data.

Applied to our networks, this method yielded a partition with 123 groups and modularity coefficient $Q \sim 0.55$ for the Twitter data; and 4 groups and modularity coefficient $Q \sim 0.42$ for the Zachary's Karate club data. For the Twitter data, most nodes (96%) are contained in the three largest communities: community 1 is formed by ∼48,000 nodes; community 2 is formed by ∼6,000 nodes; and community 3 is formed by ∼67,000 nodes. The two largest communities correspond, largely, to the two social movements considered (i.e. Occupy and 'indignados'). For the Zachary data, most nodes are contained in the two largest communities that correspond to the two factions: community 1 is formed by 12 nodes, and community 2 is formed by 11 nodes. The other two smaller communities (sizes 6 and 5) are subgraphs where some of the nodes are not directly connected to the leaders of each faction and rely on the mediation of another node. This falls in line with the insights gained from the ethnographic observations: "Not all individuals in the network were solidly members of one faction or the other. Some vacillated between the two ideological positions, and others were simply satisfied not to take sides" [8]. The algorithmic community partition identifies both the two factions and the indecisive club members. Figure 2 shows the community partition for the two networks.

We are aware that the fast greedy approach and the partition it yields may not be the most resolved or the most stable in terms of different criteria, for instance persistence [28, 29] or stability [30, 31]. Determining the stability or persistence of the method is beyond the scope of the current work. Instead, we settle for a partition that we know yields meaningful results and is coherent with a substantive interpretation of the data. The fast greedy approach is also just one of the many methods we could have employed to identify communities in the networks; however, its validity is supported by the fact that the resulting partitions are aligned with contextual information about the data: for the Zachary network, the communities identified are consistent with the ethnographic narrative of the original study [8]; for the Twitter data, the outputs reflect the divide between 'indignados' and Occupy users, and we know that different community detection methods yield

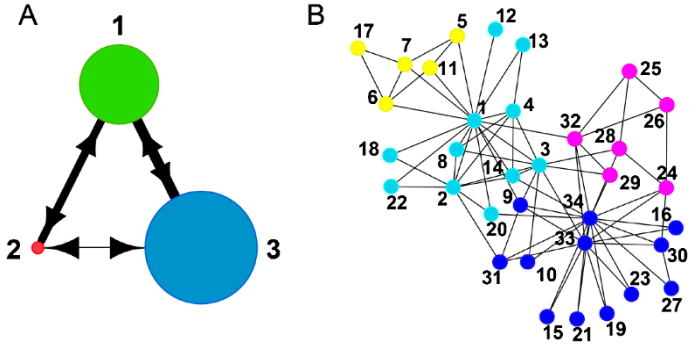


Figure 2 Community partition for the Twitter and Zachary's networks. Panel **A** shows the three largest communities of the Twitter network, and follower-following connections across them; panel **B** shows the full Zachary's network color-coded by community membership and visualized according to the Fruchterman-Reingold layout algorithm.

qualitatively similar results [24]. Because of this, we are confident that the communities identified by the fast greedy approach have empirical validity.

Finally, we consciously chose a method for the identification of non-overlapping communities because the three role schemes we test required us to do so - different theoretical goals or empirical settings might benefit from a role scheme that allows overlapping membership. However, prior research has established that disjoint communities apply important constraints to dynamics on networks, specifically to dynamics of information diffusion (see, for instance [32]). It remains an empirical question whether allowing for overlapping membership (where it makes substantive sense) can improve the modeling and explanation of those dynamics.

Once obtained, we used the community partition to determine vertex roles as defined in both the GF and the GA schemes. For the Twitter data, we first compiled the contribution matrices C^{in} and C^{out} of N nodes to M modules, where the rows of in- and out-C correspond to nodes and the columns correspond to modules. The elements of $C_{i\alpha}$ are the number of links that node i dedicates to (or receives from) module α, and can be easily obtained by multiplying the adjacency matrix of the network A_{ij} and the partition matrix S [33]. For the Zachary data, we adapted these computations to take into account that ties are undirected. These matrices summarize the tie contributions necessary to apply the GF and GA classification schemes.

The GA scheme exploits the modular information to measure the importance of a node within its community and as assessed by its inter-community bridging capability. The within-module relevance of a node is quantified through a standardized measure of the number of links the node devotes to its own community:

$$z_i = \frac{\kappa_i - \overline{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}}, \tag{2}$$

where κ_i is the number of links of node i to other nodes in the same module s_i, $\overline{\kappa}_{s_i}$ is the average of κ within module s_i, and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in module s_i.

In other words, this score measures how many standard deviations away nodes are from the mean of their communities - the extent to which they are outliers in their own groups.

The inter-modular connectivity, on the other hand, is quantified with the participation coefficient:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i} \right)^2, \tag{3}$$

where κ_{is} is the number of links of node i to nodes in modules s, and k_i is the total degree of node i. This coefficient is closer to 1 if the links are uniformly distributed among all modules and 0 if all links are within the same module.

Although the GA method has proved fruitful in the study of metabolic networks [34] and in other areas [7, 35], it is far from ideal when analyzing directed networks, where a node might have a prominent role because of its out-degree but a modest one when it comes to in-degree. In communication networks this difference matters: it amounts to differentiating sources of information from spreaders. In the context of online networks, and Twitter in particular, this difference is especially relevant since a user can follow a potentially unlimited amount of accounts, thus critically biasing the validity of the mea-

sures on which the GA roles are defined. Throughout this work we report on GA roles as obtained from the C^{out} contribution matrix. We chose to report the results for out-going links following the intuition that the role as an information producer is more relevant than the role as a receiver - admittedly an arbitrary decision, which ought to be interpreted as an additional reason for us to develop a method that integrates in- and out-connectivity in a more meaningful way.

The hybrid method for role detection we propose (HM) is based on both the GF and GA schemes. We employ two different metrics to define the two-dimensional space from where role regions are drawn. Compared to the GA method (summarized in Figure 1, panel B) we don't use node degree to define the vertical axis, but instead a local measure of brokerage base on paths of length two, similar to the notion of structural constraint [8]:

$$C_i = \sum_j \left(a_{ij} + \sum_k a_{ik} a_{kj} \right)^2. \tag{4}$$

As originally defined, the constraint of a node i can reach a maximum of k^3 and a minimum of k. A low constraint C_i indicates that node i has open triads in its personal network or, to put it differently, that it participates in paths of length two as those captured by the GF method (summarized in Figure 1, panel A). Low constraint indicates that a node has a high brokerage potential. To ease the comparison with the GA scheme, we normalize C_i so that it falls into the $[0,1]$ interval; and we invert it $(1 - C_i)$ so that higher values indicate higher brokerage scores. Also, we scale the resulting quantity by k to ensure that nodes that fill many topological gaps are placed higher in the ranking. In the end, the new score reads

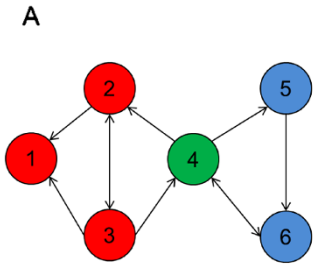
$$C_i' = k\left(1 - \frac{C_i - k_i}{k_i^3 - k_i}\right). \tag{5}$$

We define the horizontal axis of the two-dimensional space with a module-dependent measure that captures brokerage on a mesoscale. Our approach to this relies, again, on the GF scheme (Figure 1, panel A) but we depart from the original proposal in that we obtain the group partition using the network structure itself (as opposed to some exogenous node attribute). Using the in- and out-contribution matrices C_{in} and C_{out}, we measure the frequency with which a node is involved in each of the 5 role categories as defined by GF. We keep these counts in a vector g.

To project vector g onto single scalar, we weight it according to the following logic: because we are interested in the extent of extra-modular connections (which help us identify brokerage opportunities on a meso-level), we sort the GF role categories in the following order, from low to high: (1) coordinator; (2) gatekeeper; (3) representative; (4) itinerant; and (5) liaison. We then weighted each of the vector g's components by their position in the ranking, such that

$$s = 1c + 2g + 3r + 4i + 5l. \tag{6}$$

The idea behind this operationalization is that higher ranked brokers play a more important role in building global bridges, given the modular structure of the network and our choice to use modules as the group partition. A node playing the liaison role, for instance,



B

	local brokerage C'_i	global brokerage s_i
1	0.000	0.000
2	1.000	0.200
3	2.250	0.160
4	2.625	0.920
5	0.000	0.080
6	0.000	0.120

Figure 3 Local and global brokerage in an example network with three groups. This toy network illustrates our measures of local and global brokerage. Node color represents group affiliation or community. Node 4 has high local and global brokerage scores because it spans several communities and it sits on several paths of length two. Node 6 is highly constrained on the local level, but it spans a hole across communities, hence its higher global score.

bridges more structural holes at the mesolevel (i.e. it creates ties across communities) than a node playing a coordinating role (i.e. where ties span holes that are internal to the same community). Gatekeepers are lower ranked than representatives because we assume them to have a more passive role in information exchange - although arguably this depends on the actual dynamics of communication. Since we want s to lie in the interval $[0,1]$, we normalize the sum by the maximum possible count of the liaison role l in vector g, i.e. the score s^* that an ideal super-connector node would display in the network under study, i.e. $s^* = 5k_{\max}$.

Figure 3 illustrates how these two measures of brokerage operate. Panel A depicts an example network with nodes classified in three groups (coded by different colors); panel B shows the node scores for the local and global measures of brokerage. The scores indicate that node 4 is both a local and a global broker - in the sense that it controls paths linking nodes classified in different groups; node 6, on the other hand, is a broker in the global sense, but it is highly constrained on the individual level. These two measures are thus not necessarily correlated and they help identify network positions that might be functional for different reasons.

4 Findings

There are two main questions we want to answer with the methods introduced above: Do nodes that share similar network positions behave similarly in the exchange of information? And which nodes are more significant in allowing information to diffuse? To answer these questions, we first apply and compare the three methods, and then assess their definition of structural similarity through the lens of actual node behavior.

4.1 Distribution of roles

Figure 4 plots the frequency counts for each of the roles defined by the GF, the GA, and the HM methods. The upper panels summarize the Twitter data, the lower panels summarize Zachary's data (note that the gatekeeping and representative roles are indistinguishable because this network is undirected). In all three methods, we are using the same network topology and the same group partition, drawn from the modularity maximization applied to the follower network (in the Twitter case) and the face-to-face communication network (in the Zachary's data case). In all cases, the role regions are also divided using the 95%

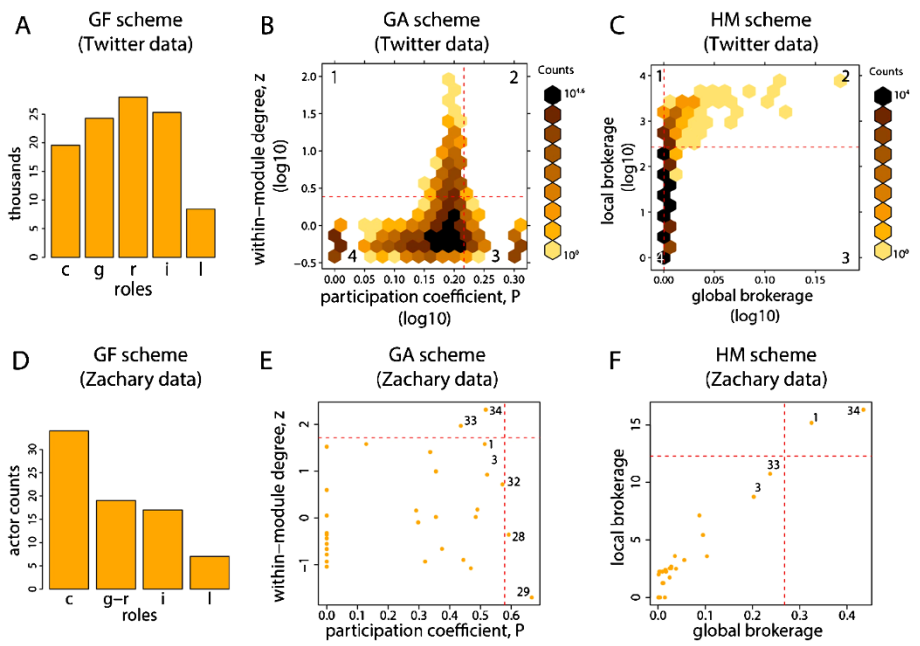


Figure 4 Distribution of roles according to the three methods. Panel **A** plots the frequency counts for the five GF roles, that is, the number of nodes that predominantly play each role. Panels **B** and **C** show the binned scatterplots mapping the distribution of nodes on the two-dimensional space defined by the degree-based method (GA) and the brokerage-based method (HM). The GA method is more restrictive in the definition of global connectors (region 2), with no nodes in that region. The lower panels summarize Zachary's data. Because this network is undirected, the gatekeeping and representative roles are indistinguishable. The HM sets two nodes clearly apart from the rest; these nodes are the leaders of the two factions identified in the original ethnographic study.

percentile of the distributions as a threshold. What changes across row panels is how roles are defined: using only local information (GF); using only global information (GA); and using both global and local information (HM). In region 1 of panel B, for instance, we have the 'provincial hubs', that is, nodes that are very well connected compared to the average of their communities but who do not connect well with other groups; in region 1 of panel C, on the other hand, we have nodes that have high brokerage scores in their local, personal networks, but who do not broker connections with nodes in other communities.

These alternative definitions of roles respond to different theoretical intuitions of why connectivity in networks matters. The first (GF) assumes that nodes that directly mediate communication between pairs that would be disconnected otherwise can benefit from that close arbitration; however, this definition does not tell us much about the larger network structure, and how central or significant brokers are in the overall picture. The second method (GA) presumes that relative centrality and the diversity of connections determine the relevance of a node: it relies on identifying the most prominent nodes within and across groups. Finally, the third method (HM) undermines the relevance of centrality and highlights, instead, the importance of mediation both at the local and global levels. These operationalizations result in substantially different classification of nodes - in the context of our data, for instance, the GA schema turns out to be very restrictive in the definition of region 2 ('connector hubs'), where no nodes can be found in either of the two datasets.

Most nodes in the Twitter network play (predominantly) a representative role (GF scheme); most are peripheral: they have a low within-module degree and low participa-

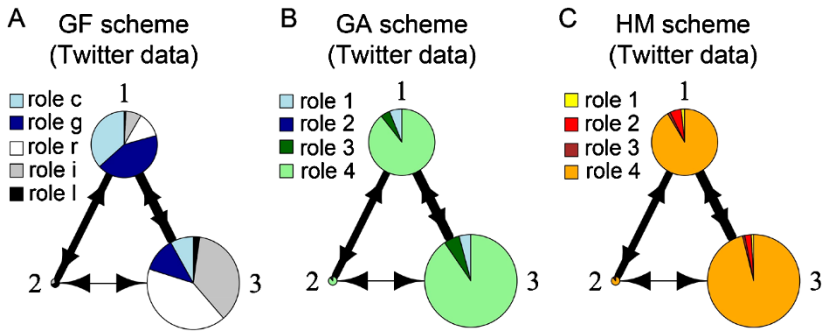


Figure 5 Distribution of roles by community in the Twitter network. Pie charts across panels correspond to the three largest communities in the following-follower Twitter network; the slices are proportional to the number of nodes playing the given role. The GA and HM schemes are more homogeneous in the distribution than the GF scheme.

tion coefficient (GA scheme); and most have low brokerage scores, both at the global and local levels (HM scheme). In the Zachary network, the HM scheme sets two nodes clearly apart from the rest - these nodes are the leaders of the two factions identified in the original ethnographic study [8], and the centers of the two largest communities identified in Figure 2, panel B. The GF method fails to identify the relevant leaders because all actors are predominantly coordinators, and it only identified a small elite of 7 actors of which the two leaders form part. The GA method, on the other hand, gives more prominence to actors that are central in the smaller communities and hides the instigators of the split. Only the HM role classification clearly sets these two pivotal actors apart.

For the Twitter data, results on this descriptive level are not that intuitive: the network is too large to be able to identify relevant identities, and we lack the ethnographic work that Zachary did. On the aggregate level, as Figure 5 shows, we can see how roles distribute by community across methods. The GF scheme, for instance, shows that community 1 (where most users communicate about Occupy) is mostly formed by coordinators and gatekeepers, whereas community 3 (with a majority of 'indignados' users) is mostly formed by representatives and itinerants. Most liaison roles are in the smaller community 2. The GA and HM schemes are, by comparison, more homogeneous: the three communities are mostly populated by peripheral users, and users with low brokerage scores.

Overall, these roles are defined on the basis of network topology only: they tell us that there are little opportunities for communication across communities, and that the vast majority of users don't control many direct diffusion channels - either because they are not well connected, or because their connections are redundant. The following section considers if this distribution of roles is associated with actual communication dynamics - which is what Zachary couldn't test with his data since his definition of ties implied face-to-face communication.

4.2 Association of roles with activity

The relevance of role schemes relies on their ability to sort out similar behavior. In the context of Twitter data, the expectation is that users with structurally similar positions will behave similarly in the exchange of information. We make use of RTs and mentions as a way to assess communication dynamics. We separate RTs from mentions because they allow information to flow in opposite directions and respond to different logics: RTs are

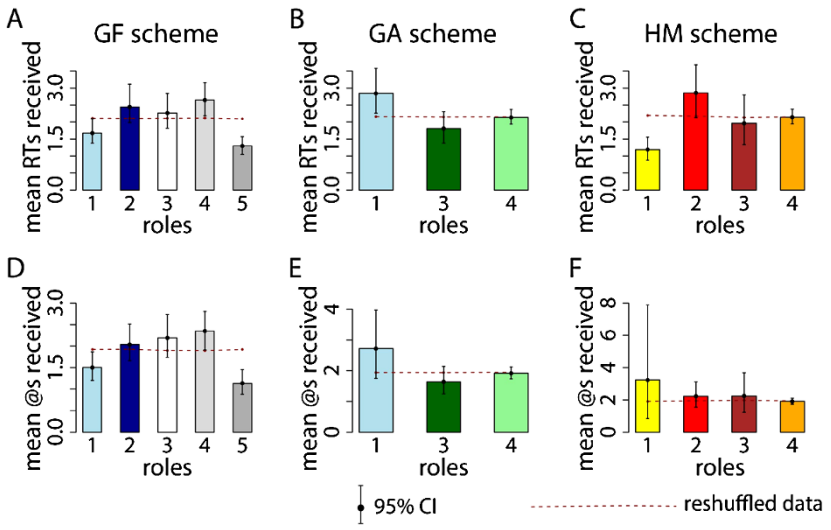


Figure 6 Authority and salience in conversation by roles. The authority of users is measured as the number of RTs received; salience is measured as the number of mentions received. The random benchmark is based on 1,000 permutations randomizing role assignment. The 95% confidence intervals are based on bootstrapped percentiles. Only the HM scheme is able to differentiate roles that are significantly associated to higher authority and salience in conversation.

about diffusing information; mentions are about raising awareness or engaging in conversation.

The bar plots in Figure 6 assess the authority and salience of users classified by roles. Authority is measured as the number of RTs received: the assumption is that a higher number of re-tweets signals higher prominence as a source of information. Salience, on the other hand, is measured as the number of mentions received: more mentions indicates that a user is better recognized and more acknowledged by other users engaging in conversation. Although the number of RTs received could also be interpreted as a measure of 'salience', we differentiate it from 'authority' to emphasize that users become prominent in the exchange of information for different reasons: RTs are about retransmission; mentions are about engaging in conversation.

To have a benchmark for comparison, we reshuffled the roles variable for each of the schemes (1,000 permutations), and calculated the mean number of RTs and mentions for the randomized data. To further assess the significance of the comparisons, we also calculated the confidence intervals for the mean values associated to each role (95% level based on bootstrapped percentiles). In panels A and D, for instance, we see that the users who predominantly engage in coordination or liaison roles are less central and visible than users acting as gatekeepers, itinerants, and - most significantly - representatives. Panels B and E suggest, on the other hand, that 'provincial hubs' (nodes with high within-module degree but low participation coefficient) are both more central and salient, although these differences are not significant.

The HM scheme uncovers differences that the other two methods fail to capture: the most re-tweeted users are those playing the role of local and global brokers; but the most salient users are brokers only on the local level (panels C and F). This suggests two things: first, that users with role 2 are essential conduits for global flows of information: they not only have more brokering ties across communities, they are also re-tweeted more often,

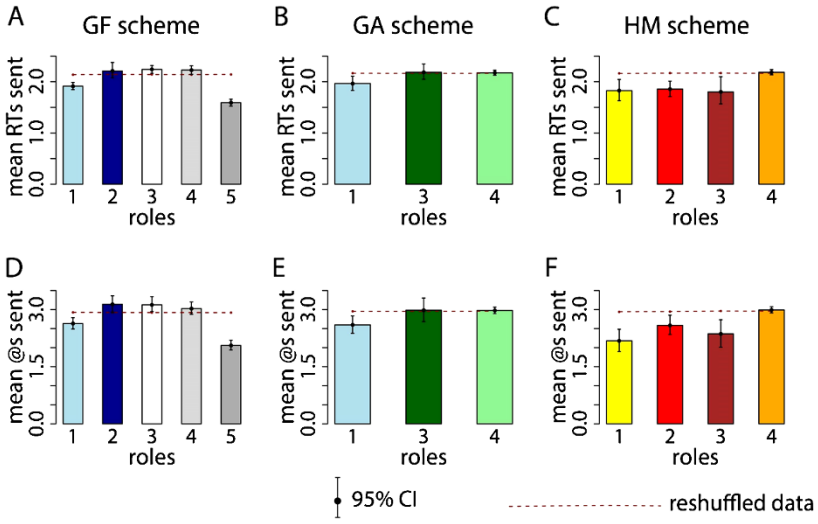


Figure 7 Retransmission and outreach efforts by roles. Retransmission efforts are measured as number of RTs made; outreach is measured as number of mentions made. The HM scheme sets role 4 apart as the source of most of the communication efforts.

which is a precondition for information cascades; and two, the findings also reveal that there is an organic division of labor in the network: users acting as authorities are different (and have different network positions) than users that become the target of conversation more often. The HM scheme identifies this division of labor more clearly than the other two methods.

Figure 7 provides additional evidence on how roles distribute in the network, this time tracking retransmission and outreach efforts (measured as the number of RTs and mentions made, as opposed to received). Activity levels according to the GF scheme do not vary much compared to those depicted in Figure 6; this suggests that this scheme is not very useful to distinguish nodes that are relevant because they receive attention versus those that are relevant because they produce the attention. Both the GA and HM schemes suggest that most of the messages are generated by peripheral roles (i.e. the less significant nodes in terms of authority and salience). Only the HM method, however, clearly sets apart the group of nodes that produce most of the RTs and mentions: they are the users with low brokerage scores both at the local and global levels.

5 Discussion

Role schemes aim to identify network features that make nodes structurally similar. The assumption is that nodes that are similar in their connections will exhibit a similar behavior or fulfill a comparable function within the system the network maps. The three methods we compare here use the modular structure of the network to define roles, but they differ on how they capture the heterogeneity of individual positions: the GF approach makes use of local features (i.e. mediation in paths on length two); the GA approach focuses on the distribution of degree centrality, and the number of ties that link to different communities, measures both that require global knowledge of the network; and the HM approach, which combines the previous two by assessing brokerage at the local (i.e. personal network) and global (i.e. inter-community ties) levels.

This focus on different network features results in alternative role classifications that, as the findings reported show, capture with more or less success differences in behavior. On the basis of our data, we can conclude that the hybrid model combining local and global measures of brokerage is able to better characterize significant behavior. In the Zachary network, it helps identify the leaders of the two opposing factions; in the context of the Twitter network, it helps identify the division of labor that emerges in the orchestration of political action: on the one hand, users playing role 2 (local and global brokers) act as sources of information, but the most visible users play role 1 (they are the local brokers); on the other hand, most of the messages sending those RTs and mentions are generated by users playing role 4. This allocation of roles not only helps uncover an organic division of labor (organic for unplanned); it also allows inferring the logic that lies behind communication dynamics. If RTs are used to disseminate information from a source to an audience, mentions are used to try to gain the attention of other users or engage in direct conversation. It is not coincidental that users receiving many RTs have more global networks than users receiving many mentions.

There are many other alternative schemes that could possibly shed light on communication behavior; and many of the parameters we fixed in our operationalizations could also be modified. For instance, applying an alternative community detection method would have yielded a different partition; or we could have employed a partition completely independent of network topology, based on some node attribute like membership or affiliation to specific organizations. We think, however, that the modular structure of networks contains significant information about how social systems self-organize, and so we advocate for the topological approach to the classification of nodes (unlike the GF scheme as originally formulated). The two datasets we use exhibit a clear modular structure that responds to identifiable social factors (i.e. conflict dynamics, membership to social movements). This makes the use of our partition substantively meaningful - it reflects well the groups that operate within the network - but the method might yield uninterpretable results if the forces driving the formation of modules in networks are unknown.

Thus, before generalizing our approach and making it applicable to other settings, careful consideration should be given to how the partition is generated and whether it makes sense for the empirical data under analysis. A relevant aspect to take into account in this respect is that in some scenarios overlapping communities might be more meaningful. The three schemes we consider here are based on the assumption that communities are exclusive, but future research should address the effects that multiple memberships would have on the definition of roles.

The second parameter that could be modified is the threshold we use to delineate role regions. We used the 95% percentile, which is arguably an *ad hoc* heuristic. This threshold was chosen to give prominence to the tail of distributions that are skewed, but future work should also assess changes in the performance of role schemes with shifts in this threshold. The sensitivity of role classification to different heuristics is, ultimately, an empirical question: for some data structures, some divisions will yield more informative classifications than others.

On a theoretical level, we argue that a role scheme based on brokerage is more discriminating than a scheme based on degree. The findings show that roles based on brokerage result in a reduction of the network that is more informative about the functions that nodes play. The reason, we sustain, is that the analysis of brokerage allows integrating more lev-

els of analysis (from local bridges in open triads to ties linking modules), and it thus offers a more informative definition of structural similarity.

6 Conclusion

This paper extends previous sociological research on the importance of brokerage to the analysis of large networks of communication. Our goal was twofold: to determine whether users that share similar network positions behave similarly; and to identify the roles that are more significant for information diffusion. Our findings confirm the expectation that nodes creating bridges within and across communities (role 2 in the HM scheme) are more likely to be used as sources of information: they are the leaders in the Zachary network and the users receiving a higher number of RTs in the Twitter network. This suggests that these users are crucial in the global exchange of information. For the Twitter data, we also found that roles reveal a division of labor in how communication takes place: users that act as authorities are not the most visible; and the most central or visible are not the most active. Overall, the relative performance of different role schemes depends on the empirical data; but the methods to define structural similarity can be generalized beyond the specificities of particular data sets. Improving our definitions of structural similarity is important because they offer a method to model and simplify complex structures - an exercise that is increasingly necessary to make sense of large-scale digital networks and the communication they facilitate.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SGB and JGH designed research, performed analyses, and wrote the paper. NW collected the data and performed preliminary analyses.

Author details
[1]Annenberg School for Communication, University of Pennsylvania, Philadelphia, USA. [2]Oxford Internet Institute, University of Oxford, Oxford, UK. [3]Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar.

Received: 10 April 2014 Accepted: 23 October 2014 Published online: 07 November 2014

References

1. Rainie L, Wellman B (2012) Networked. The new social operating system. MIT Press, Cambridge
2. Newman MEJ (2012) Communities, modules and large-scale structure in networks. Nat Phys 8:25-31
3. Serrano MÁ, Boguñá M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA 106:6483-6488
4. Harrison CW, Boorman SA, Breiger RL (1976) Social structure from multiple networks. I. Blockmodels of roles and positions. Am J Sociol 81:730-780
5. Boorman SA, Harrison CW (1976) Social structure from multiple networks. II. Role structures. Am J Sociol 81:1384-1446
6. Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1:49-80
7. Guimera R, Nunes Amaral LA (2005) Functional cartography of complex metabolic networks. Nature 433:895-900
8. Zachary WW (1977) An information flow model for conflict and fission in small groups. J Anthropol Res 33:452-473
9. Burt RS (2005) Brokerage and closure. An introduction to social capital. Oxford University Press, Oxford
10. Burt RS (1992) Structural holes. The social structure of competition. Harvard University Press, Cambridge
11. Gould RV, Fernandez RM (1989) Structures of mediation: a formal approach to brokerage in transaction networks. Sociol Method 19:89-126
12. Gould RV (1989) Power and social structure in community elites. Soc Forces 68:531-552
13. Baños R, Borge-Holthoefer J, Wang N, Moreno Y, González-Bailón S (2013) Diffusion dynamics with changing network composition. Entropy 15:4553-4568
14. González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Sci Rep 1:197
15. Conover MD, Davis C, Ferrara E, McKelvey K, Menczer F, Flammini A (2013) The geospatial characteristics of a social movement communication network. PLoS ONE 8:e55957
16. Conover MD, Ferrara E, Menczer F, Flammini A (2013) The digital evolution of Occupy Wall Street. PLoS ONE 8:e64679
17. Merton RK (1957) Social theory and social structure. Free Press, New York

18. Burt RS (1982) Toward a structural theory of action: network models of social structure, perception and action. Academic Press, New York
19. Burt RS (1976) Positions in networks. Soc Forces 55:93-122
20. Burt RS (1978) Cohesion versus structural equivalence as a basis for network subgroups. Sociol Methods Res 7:189-212
21. Aral S, Van Alstyne M (2011) The diversity-bandwidth trade-off. Am J Sociol 117:90-171
22. Spiro ES, Acton RM, Butts CT (2013) Extended structures of mediation: re-examining brokerage in dynamic networks. Soc Netw 35:130-143
23. Burt RS (2004) Structural holes and good ideas. Am J Sociol 110:349-399
24. González-Bailón S, Wang N (2013) Networked Discontent: The Anatomy of Protest Campaigns in Social Media SSRN: http://ssrn.com/abstract=2268165
25. González-Bailón S, Wang N, Rivero A, Borge-Holthoefer J, Moreno Y (2014) Assessing the bias in samples of large online networks. Soc Netw 38:16-27
26. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104
27. Csárdi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal:1965
28. Arenas A, Fernández A, Gómez S (2008) Analysis of the structure of complex networks at different resolution levels. New J Phys 10:053039
29. Arenas A, Fernández A, Gómez S (2008) A complex network approach to the determination of functional groups in the neural system of *C. elegans*. In: Liò P, Yoneki E, Crowcroft J, Verma D (eds) Bio-inspired computing and communication, vol 5151. Springer, Berlin, pp 9-18
30. Le Martelot E, Hankin C (2012) Multi-scale community detection using stability optimisation. Int J Web Based Communities 9:323-348
31. Le Martelot E, Hankin C (2013) Fast multi-scale detection of relevant communities in large-scale networks. Comput J 56:1136-1150
32. Weng L, Menczer F, Ahn Y-Y (2013) Virality prediction and community structure in social networks. Sci Rep 3:2522
33. Arenas A, Borge-Holthoefer J, Gómez S, Zamora-López G (2010) Optimal map of the modular structure of complex networks. New J Phys 12:053009
34. Baños R, Borge-Holthoefer J, Moreno Y (2013) The role of hidden influentials in the diffusion of online information cascades. EPJ Data Sci 2:6
35. Olesen JM, Bascompte J, Dupont YL, Jordano P (2007) The modularity of pollination networks. Proc Natl Acad Sci USA 104:19891-19896

doi:10.1140/epjds/s13688-014-0032-y
Cite this article as: González-Bailón et al.: **The emergence of roles in large-scale networks of communication.** *EPJ Data Science* 2014 **3**:32.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com