Noname manuscript No.
(will be inserted by the editor)

A new approach to role and position detection in networks

Davide Vega · Matteo Magnani · Danilo Montesi · Roc Meseguer · Felix Freitag

Received: date / Accepted: date

Abstract We rethink and extend the concepts of position and role in a network, basing them on various well-known measures that were not previously associated to these concepts, like geodesic distance and modularity. The effectiveness of our new role and position detection algorithms is evaluated both qualitatively and quantitatively, on synthetic and real data, showing that we can identify new types of meaningful patterns in networks.

Keywords Position · Role · Detection · Networks · Communities · Betweenness · Shortest path · Blockmodeling · Approximation

1 Introduction

Structural analysis on networks intends to capture and interpret how nodes are related to each other according to the network topology. When applied to social networks, structural analysis is able to identify key actors[1] or groups of actors whose connectivity influences the dynamics of the system. Three typical ways of grouping actors based on their connections consist in identifying *communities*, *positions* and *roles*. While related, these are three distinct types of groups and they typically require distinct algorithmic treatments.

D. Vega · D. Montesi
University of Bologna
E-mail: {davide.vegadaurelio,danilo.montesi}@unibo.it

M. Magnani
Uppsala University
E-mail: matteo.magnani@it.uu.se

R. Meseguer · F. Freitag Universitat Politecnica de Catalunya
E-mail: {meseguer,felix}@ac.upc.edu

[1] Individuals or organizations corresponding to the nodes in the social graph.

1.1 Communities, positions and roles

To briefly recall the difference between community, position and role, we use Padgett's social network representing business relationships among Florentine families during Renaissance [Breiger and Pattison(1986)] (Fig. 1).

A **community** is a cohesive group of actors, with many connections inside the group and fewer relationships with other actors outside it. As an example, the five nodes on the top of the figure form a community. Each of the families inside this community has from two to four ties among their members, and no more than one tie with other actors in the network. Several methods have been developed along the years to find different kinds of communities, e.g., overlapping and hierarchical, keeping in their core the basic idea of cohesiveness [Fortunato(2010), Coscia et al(2011)Coscia, Giannotti, and Pedreschi].

Positions, instead, are based on the concept of interchangeability: two actors in the same position can be swapped without changing their relationships with other actors in the network. In our example, the three families *Salvati*, *Pazzi* and *Tornabuon* are in the same position, because they are all connected to the *Medici* family and none of them is connected to any other family. Hence, these three actors are interchangeable: swapping the connections of two of them would not change the topology of the network. Differently from community detection, in general it is not important for actors in the same position to be connected to each other.

Finally, the concept of **role** refers to actors with similar patterns of connectivity, independently of the specific actors to whom they are connected. In our running example, the *Barbadori* family has the role of connecting two otherwise disjoint parts of the network. From this point of view, it does not matter who exactly is connected to them: if the *Barbadori* family were connected to *Salviati* instead of *Ginori*, they would still play the same role in the network, but from a different position.

In summary, nodes in the same position or role are similar according to their relationships — similarity — with the other nodes. Therefore, using different kinds of relationships and thus different similarity measures we can define and identify completely different social structures.

1.2 An extended model

In this work we focus on position and role analysis. To understand our contribution we can consider a standard mathematical definition of the concept of position. In the model based on the so-called *structural equivalence*, two actors are in the same position if and only if they are connected with the exact same actors. Let $G = (V, E)$ be a graph representing a social network, where V is a set of nodes representing actors and $E(i, j) = 1$ if nodes i and j are connected, 0 otherwise. Then, we can say that two nodes i and j are structurally equivalent (and so in the same position) if and only if [White and Reitz(1983)]:

$$E(i, k) = E(j, k) \quad \forall k \in V; k \neq i, j \tag{1}$$

In our previous example, *Salvati* and *Pazzi* are both connected to *Medici* (so, $E(\textit{Salvati}, \textit{Medici}) = E(\textit{Pazzi}, \textit{Medici}) = 1$), and for every other actor k they are not connected to it, so E equals 0 for both families in all other cases.

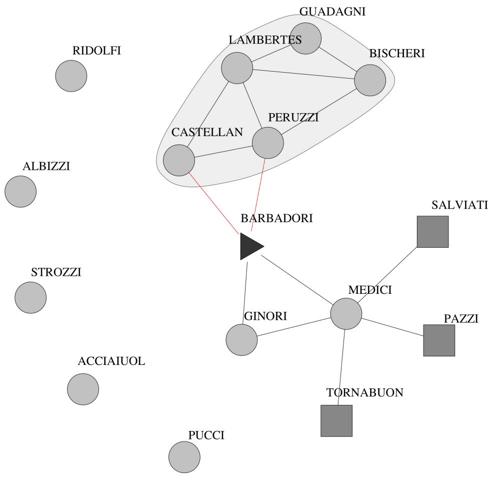


Fig. 1: Padgett's business family network, where we have highlighted a community (gray area), a structurally equivalent position (square nodes), and a bridging role (triangular node)

This basic definition has been extended in many ways in the literature — e.g., replacing E with other comparison measures (that we generically notate as D in the following) or using this formula in an iterative process, as reported in Section 5 [Doreian(1988)]. However all existing variations of this definition rely on comparing a pair of actors against *one single item at a time*, here represented by the symbol k. In this paper we extend this perspective on position analysis by replacing Eq. 1 (and its generalizations) with a set-based definition. Two nodes i and j are, according to our model, in the same position if and only if:

$$D(i, S_k) = D(j, S_k) \quad \forall S_k \subseteq V \tag{2}$$

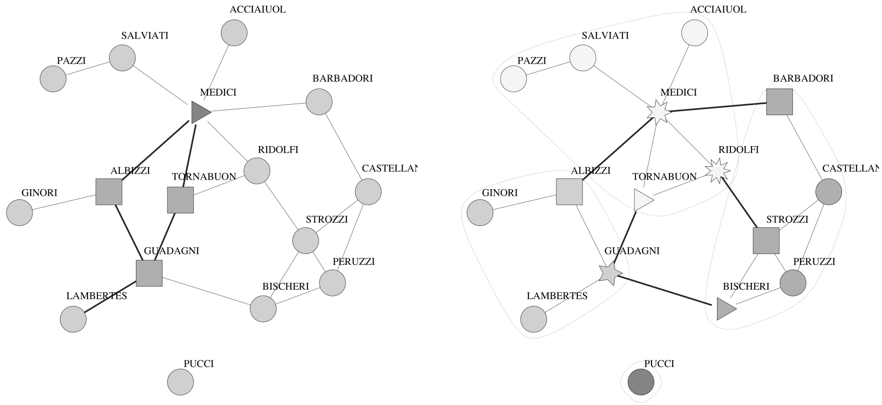
The main formal difference, whose significant implications will be explored in our experimental evaluation, is the usage of a set S_k instead of the singleton k. This allows us to express roles and positions using several more similarity measures — here generically notated as D — to compare actors, that would not make sense for single pairs of nodes and have thus been overlooked in the literature on position detection.

As a concrete example consider Fig. 2, showing two representations of the Padgett's marriage network where the shape of the nodes represent their position. We can observe that families *Albizzi*, *Guadagni*, *Barbadori* and *Strozzi* are connected to totally different nodes here, that are themselves in different positions. So, they would not be considered being part of the same position by existing methods. For example, if we check the *Medici* family, *Albizzi* and *Barbadori* are connected to it while *Guadagni* and *Strozzi* are not. If we check *Lambertes* only *Guadagni* is connected to it, in the same way as only *Strozzi* is connected to *Peruzzi*.

However, if we now consider the pair $\{Lambertes, Medici\}$ and a comparison function:

$$D(i, \{k, q\}) = \begin{cases} 1 \text{ if } i \in \text{shortest path betw. k and q} \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{3}$$

we can see in Fig. 2a that both, *Albizzi* and *Guadagni*, are on a shortest path between them — noted with thicker edges. If we check other pairs of nodes, we can see that this is true in several other cases, e.g., to efficiently go from *Bischeri* to *Ginori* we should also pass through *Albizzi* and *Guadagni*. In summary, *Albizzi* and *Guadagni* are included in the same position because they share the same relationship with other *pairs of nodes*: to go from specific parts of the networks to other specific locations we can indistinguishably pass through any of these two nodes.



(a) Positions based on shortest paths

(b) Positions based on ties to communities

Fig. 2: Padgett's marriage family network and approximate positions defined as: (2a) being part of the shortest path connecting pairs of nodes, (2b) being a bridge between the same communities (the concept of approximation will be defined later in the paper).

If we now focus on Fig. 2b we can see a different positional analysis based on the comparison function:

$$D(i, S_k) = \begin{cases} 1 \text{ if } i \notin S_k \text{ and } \exists k \in S_k, \{i, k\} \in E \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{4}$$

where each set S_k corresponds to one of the four communities in the graph. According to this new similarity measure, *Albizzi*, *Barbadori* and *Strozzi* are now in the same position as they are all connected to some members of the community on the top (*Medici* or *Ridolfi*) and to members of their own community, but not to other communities in the network. *Guadagni*, instead, is connected to two different communities apart from its own — through *Tornabuon* and *Bischeri*. As a result, *Albizzi*, *Barbadori* and *Strozzi* are in the same position because they are acting as gateways only between their own community and the largest one. *Guadagni* is in a different position.

These examples highlight three important aspects of our work. First, using this set-based comparison and the corresponding similarity functions we are identifying **new types of position**. Specifically, the first example identifies actors that can be inter-changed without significantly affecting the messages that would pass through them. As it appears from Fig. 2a, nodes in the center and nodes at the periphery of the network tend to be included into different positions — this being one of the features distinguishing the new type of positions from the ones we can obtain using existing methods based on pairwise comparisons.

Second, this extension allows us to use several **other network measures** involving comparisons with larger and heterogeneous sets of actors, like triangles, cliques and communities. The second example (Fig. 2b), identifies nodes controlling how information can spread across cohesive groups of nodes. Nodes connected only within their community appear all in the same position as they are not helping to spread information outside it, while nodes on the edge of different communities are identified in several different positions based on from and to which communities they can directly spread information.

Third, we obtain a clearer understanding of the often ambiguous **relationship between positions and roles**: with this extended definition, we can say that for each type of position we have a corresponding role, which considers the pattern of connectivity but not the specific individuals involved. In our first example (Fig. 2a), the *Ridolfi* and *Strozzi* families might also constitute a position, different from the previous one — e.g., they are in the shortest path between *Peruzzi* and *Medici*, which is not the case for *Albizzi*. However, the number of shortest paths traversing them (that is, their betweenness) is similar, making them play a similar role in the network. In Fig. 2b the *Tornabuon* and *Ridolfi* families play the same role as *Albizzi*, *Barbadori* and *Strozzi*, as all five are bridging two communities; the specific communities that are connected determine the different corresponding positions in the network. *Guadagni* and *Medici* play a different role connecting three distinct communities. In this illustrative example the difference between these roles is small (that is, bridging two instead of three communities), but as we will see in the experimental validation of the method clearly distinct groups can emerge in larger networks.

We will mathematically formalize the relationship between role and position later in the paper, but it is worth noticing that to the best of our knowledge concepts like betweenness and community had never been formally related to the concept of position before. These and other connections directly emerge from our extended formalization once Eq. 3, Eq. 8 and other similarity functions introduced later in the paper are used.

As a future-looking note, our extended model also gives us the flexibility to study social roles and positions in other network models like **hypergraphs**, which can be represented as hyperlink adjacency matrices without losing information about the hyperlinks — as it would happen if the regular adjacency matrix is used. Another example are **multiplex/multi-relational** graphs, where our model enables the usage of measures based on paths traversing multiple types of relational ties.

1.3 Contributions

The contributions made in this work can be summarized as follows:

- We propose a novel framework for identifying new types of social positions and roles based on the structural similarities between actors and groups of actors in the network (Section 2). We noticed that traditional methods use pairwise comparisons between actors to find similar structural patterns, limiting the range of similar structures identificable. In this work we extended and adapted the current blockmodeling methods, rather than substitute them, with the purpose of finding other types of positions and roles which have been overlooked in the past.
- In Section 2.1 we present in detail two specific approaches, based on two different types of similariy measures, which are a good illustration of the flexibility of the new framework. The first group is a good example of how to redefine the traditional similarities by introducing sets of actors as a comparison unit — like positions based on ties between communities (Fig. 2b), while the second one uses the set of actors to measure new similarities that would not make sense for single pairs of nodes — like positions based on shortest paths (Fig. 2a).
- In the process of constructing the new framework, we have decoupled the several definitions of equivalence and similarity from the mathematical procedures to increase the flexibility of the framework. Therefore, given a particular structural pattern (aka similarity measure) our framework is able to compute both the associated positions (Section 2.2) as well as the equivalent roles (Section 2.3).
- Finally, we perform an original and extensive experimental evaluation based on several synthetic and real datasets, obtaining both qualitative (Section 3) and quantitative (Section 4) insights characterizing the results of the proposed method. These can be used as guidelines to decide when to apply our approach and to interpret the obtained results.

This article is an extended version of a conference paper focusing on some aspects of the method [Vega et al(2015)Vega, Magnani, Meseguer, and Freitag]. With respect to this previous work, apart from a more complete presentation and formal development, we study more types of similarity functions, introducing new kinds of positions, we provide a detailed comparison of the relationships between the concepts of role and position, and we perform an original and extensive experimental evaluation based on several synthetic and real datasets.

2 A new blockmodeling framework

Blockmodeling [Wasserman(1994)] is, to our knowledge, the most used and explored technique to detect roles and positions in social networks and, more generally, in any system that can be modeled mathematically using a graph. In blockmodeling, actors are grouped into positions — also called blocks, sometimes roles

— based on a similarity or dissimilarity measure between them. To compute this measure, actors are compared based on their social behaviour and structural connectivity in the network. In its original form, the similarity measure corresponds to the correlation between columns in the graph adjacency matrix, which results in including actors connected to the same other actors into the same position — as for the three square nodes in Fig. 1.

In this section we describe our framework for group relations which allows us to apply blockmodeling to find social roles and positions without being constrained by pairwise comparisons. The framework has two basic components: the *extended comparison function* for group measures and the *computing algorithm* for identifying positions and roles.

The extended comparison function is a two-dimensional matrix (M) that stores the similarity or dissimilarity between actors (rows) and sets of actors (columns). The algorithm used in our experiments is a generalization of the REGE/A algorithm proposed in [Borgatti and Everett(1993)] for regular equivalence.

This practical decision allows us to both, describe more clearly the differences between our framework and other traditional blockmodeling techniques and to provide a fairer comparative analysis between the different options (See Section 4). In general, any clustering algorithm already used in unsupervised blockmodeling analysis could be used instead.

2.1 Extended comparison matrix

In order to compare actors with subsets of actors we replace the adjacency matrix with a larger structure, capable of storing extended relations. The new structure is a two-dimensional matrix with $|V|$ rows — each one representing one of the original actors in the network — and with up to $2^{|V|}$ columns — representing each of the possible groups of interest. In practice, only a small fraction of these columns is necessary, depending on the chosen comparison function. In summary, the cells of the matrix contain the value of a generic function:

$$D : (V, S) \to \mathbb{R} \tag{5}$$

defined over a graph $G = (V, E)$, where $S \subseteq 2^{|V|}$. Hence, we can define the extended matrix M as:

$$M(i, S_j) = D(i, S_j) \tag{6}$$

Back to our example on detecting positions as being part of the shortest path that connects pairs of nodes, we had defined the comparison function in Eq. 3 as:

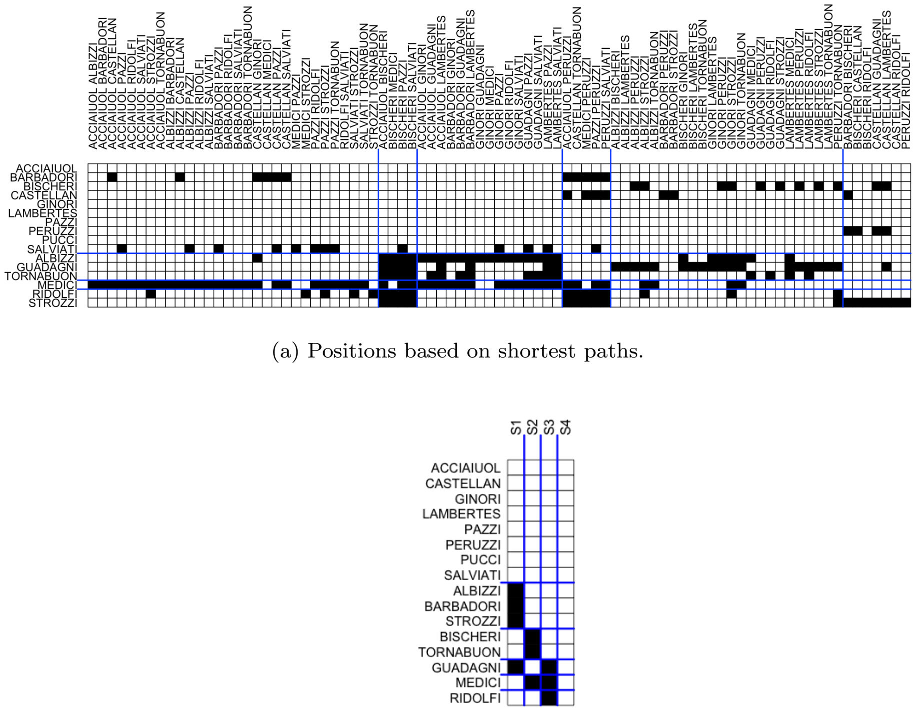
$$D(i, \{k, q\}) = \begin{cases} 1 \text{ if } i \in \text{shortest path betw. k and q} \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise} \end{cases} \tag{7}$$

Therefore, in this case S consists of all subsets of V of cardinality 2. Instead, in the example on detecting positions as being equivalently connected to other communities, the comparison function is slightly more complex, but the extended matrix is typically significantly smaller. In fact, it will contain only one column per community. Given a set of subsets of V, representing communities:

$$D(i, S_k) = \begin{cases} 1 \text{ if } i \notin S_k \text{ and } \exists k \in S_k, \{i,k\} \in E \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{8}$$

Each of the elements in S may have a different cardinality. Hence, there is no constraint regarding the size, composition or diversity of the subsets of the extended matrix M. Our extended matrix is a generalization of the regular adjacency matrix used in traditional blockmodels, where $S = \{\{u,v\}\,|u,v \in V\}$.

Fig. 3 shows the corresponding extended matrix M for both measurements. Each cell in both matrices corresponds to one binary relation between an actor i and a set of other actors: a pair $S_j = \{k, q\}$ for the positions based on the shortest path (Fig. 3a), and a group of actors $S_j = \{v_{j1}, \cdots, v_{jk}\}$ representing a cohesive community in the second case (Fig. 3b). The rows and columns have been arranged in order to group together similar positions.



(a) Positions based on shortest paths.

(b) Positions based on ties between communities.

Fig. 3: Extended matrix for the Padgett's marriage family network. Positions are defined as: (3a) being part of the shortest path connecting pairs of nodes and (3b) being a bridge between the same external communities.

The extended comparison matrix, which associates each actor in the network with the subsets of actors used to measure the equivalence, can then be used for both partitioning the set of actors into $\beta_M = \beta_1, \beta_2, \ldots \beta_m$ similar positions, or into $\rho_Z = \rho_1, \rho_2, \rho_3, \ldots \rho_z$ similar roles. In traditional blockmodeling only one of the two partitions is calculated because both — roles and positions — are considered interchangeable concepts determined uniquely by the definition of equivalence used. However, in our framework they are different concepts that can be measured from the extended matrix. Back to our examples, the extended matrix based on shortest path (Fig. 3a) can be used to detect either "actors being in exactly the same shortest paths connecting the same pairs of nodes" (positions) or "actors being in the same number of shortest paths" (roles). This definition of role is not new: it is based on the concept of betweenness. Therefore, we can see this new type of position enabled by our framework as a location-aware counter part of betweenness; where we do not just count the shortest paths but we also consider where they appear in the network. In the same way, the extended matrix based on ties between communities (Fig. 3b) can be used to detect "actors facilitating the exchange of information across the same communities" (positions) or "actors bridging the same number of communities" (roles).

2.2 Position assignment

Positions are computed by clustering the rows of the extended matrix and forming groups of actors whose relations with the same subsets are similar. Likewise, we could also compute the similarity between sets of actors — columns of the extended matrix — to know who are the influencing groups for each position. Including into the same position different nodes whose connectivity patterns are not exactly the same, but very close, is typical also in traditional blockmodeling, and is done so that small random variations in the network or small amounts of missing data do not prevent grouping together otherwise interchangeable nodes. This is discussed in more detail later.

In the literature there are many clustering algorithms that can be used to make this assignment. In order to simplify the results, and for comparison purposes with other indirect blockmodeling methods, we decided to use agglomerative hierarchical clustering. In concrete, we first generate a similarity matrix by computing the Euclidean distance between rows of our extended matrix M and then we generate the agglomerative hierarchical clustering using Ward's method [Murtagh and Legendre(2014)], whose objective function tries to minimize the total within-cluster variance at each step of the recursive algorithm. For these calculations we use the *fastcluster library* developed by Müller [Müllner(2013)] which, given a similarity matrix of $|N|$ elements, computes the resulting hierarchical clustering in $\Theta(N^2)$. Therefore, the whole positional assignment depends quadratically on the number of actors $|V|$ (clustering) and linearly on the maximum between the number of actors and the number of actors' sets $|S|$ (Euclidean distance).

In practice, the number of sets $|S|$ used by most of the common *extended comparison functions* is similar to the number of actors in the network. As an example, computing distance-based similarities between actors not directly connected requires less than V^2 sets (See Figure 3a), while computing similarities based on actors' connectivity to different communities require even fewer sets (e.g., $|S| = 4$ in a network with 16 actors, according the example presented in

Figure 3b). While in theory, there is a mathematical worst-case scenario ω where the number of sets equals all possible sub-setting of actors $|S| = S^{|V|}$, we cannot imagine a real extended comparison function using such amount of comparisons. In summary, the expected running time of the position assignment is $\Theta(V^2) + \Theta(max(V, S)) \sim \Theta(N^2)$.

Finding the optimal number of clusters (i.e., positions) is discussed in detail in Section 2.4. As examples of positional analysis, we have indicated the result of this process in Fig. 2, where blue horizontal lines separate the actors in, respectively, 4 and 6 positions.

2.3 Role assignment

The procedure for identifying roles uses the same extended matrix M and removes the association between actors and groups, so that patterns independent of the specific location in the network can emerge. In other words, for each actor i we can consider the distribution of the values in the corresponding row of the matrix, discarding the order of the elements. Mathematically, this means that we can consider each actor as a random variable x and the elements on its row as values covered by χ. Therefore, we can compare actors based on the similarity of the probability distribution associated to the corresponding random variables.

The framework does not enforce any specific method for comparing the different probability distributions, allowing to choose the most suitable one based on the input network and the extended comparison function used fo create the matrix M. In practice, most of the tests based on non-parametric models can be used. As an example, a straightforward method will be to compute the histogram of values in each row of the matrix, and then use the Pearson correlation or Phi coefficients to compare their distributions. Knowing or inferring, instead, some information about the probability distribution will allow us to apply more suitable methods, like for example the Kolmogorov-Smirnov nonparametric test [Durbin(1973)] for normal distributions.

Back to our working approaches, as our matrices are binary we can summarize the distribution by just summing up the values in each row, this is, counting number of ones:

$$A(i) = \sum_{j \in S} M(i, j) \tag{9}$$

This extra function A, which does not exist in previous blockmodeling approaches, is the one allowing us to relate positions with their complementary roles.

Table 1 shows the corresponding roles detected applying Eq. 9 to both extended matrices. Notice that for each similarity measure the same actors can be assigned to different roles. While the *Albizzi* and *Guadagni* families are assigned to the same role (ρ_4) based on the shortest-path measurement, they are identified in different roles (ρ'_2 and ρ'_3 respectively) with respect to their connectivity to different communities.

As a last note, observe that the assignment of actors into roles is not the same as their assignment into positions. While the *Medici* family, as an example, is the single member of a role and a position — using the shortest-path based similarity — *Strozzi* and *Tornabuon*, who play the same social role, are in fact in

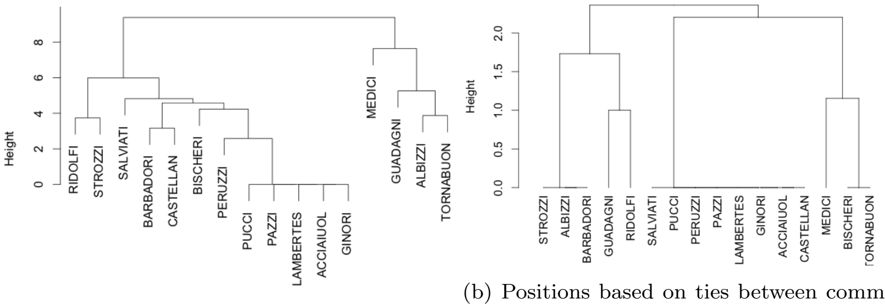
Table 1: Roles identified as being part of the shortest path between pairs of nodes or connecting the same number of communities in Padgett's marriage families network. In the bottom line, the number of 1s in the rows corresponding to the families in each role

Roles	**Shortest-path based**					**Community based**		
	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ'_1	ρ'_2	ρ'_3
ACCIAIUOL	0					0		
GINORI	0					0		
LAMBERTES	0					0		
PAZZI	0					0		
PUCCI	0					0		
CASTELLAN		7				0		
PERUZZI		4				0		
BARBADORI			11				1	
BISCHERI			11				1	
SALVIATI			13				1	
TORNABUON			16				1	
RIDOLFI			17				1	
STROZZI			17				1	
ALBIZZI				26			1	
GUADAGNI				27				2
MEDICI					50			2
	0	**4, 7**	**11-17**	**26,27**	**50**	**0**	**1**	**2**

different positions — because they are in the same number of the shortest paths, but between different sets of actors. Under a strict check of equivalences between the rows of the extended matrix, positions would be possibly finer partitions of the roles.

2.4 Approximate positions and roles

We have previously mentioned the possibility of using different degrees of freedom for each definition of equivalence, especially when the notion of structural equivalence is used. This is common practice in the blockmodeling literature as it is unlikely to find any meaningful structurally equivalent positions in networks with dense structures, like Padgett's marriage families network (Fig. 2): the normal variability in connectivity prevents us from finding two nodes with many connections and connected with the exact same other nodes. As a result, every single actor in the network will be placed in a different position with only one member. To relax the definition, and find meaningful positions and roles using indirect approaches, it is sometimes necessary to utilize some knowledge about the social network under analysis, e.g., specifying the number of expected positions.



(a) Positions based on shortest paths.

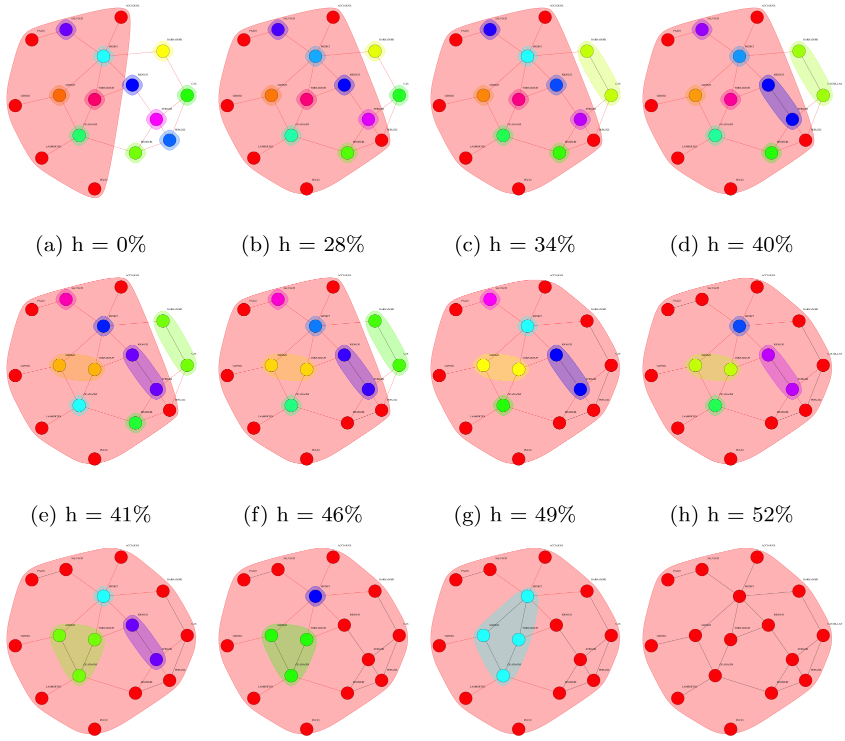
(b) Positions based on ties between communities.

Fig. 4: Dendrograms representing the hierarchical clustering of the rows of the matrix in Fig. 3 based on actor similarity

As an alternative, we can measure how good different partitions are by measuring the relative distance between the maximum height of the hierarchical clustering and the height of the cutting point in the dendrogram. We remind the reader that hierarchical clustering methods generate a dendrogram indicating how dissimilar different groups are when the agglomerative algorithm decides to merge them into a single group. In Fig. 4 we show it for our working example. Ideally, cutting the dendrogram at height $h = 0$ we look for positions that are totally indistinguishable compared with the single actors — which is equivalent to group together strictly equivalent nodes. As we pull up the cutting point we relax this constraint and we can start grouping actors into fewer positions. At maximal height all actors would be included in the same position. This relaxation of the block formation can be applied to all extended measures, and makes them comparable in terms of precision of the results.

The probability of having more or less precise positions and/or roles is also influenced by the comparison function D and the number of groups S. While the positional analysis based on communities generates a dendrogram with clear positions and roles (See Fig. 4b), the positional analysis based on shortest paths generates a more complex dendrogram (See Fig. 4a), mainly because the number of sets S considered — pairs of nodes — is significantly larger than the number of communities in the original network.

In our experiments we refer to *minimum height* (h_{min})' as the minimum relative value of h needed, for a particular clustering, to find at least one position with more than one actor. Having a higher or lower h_{min} does not imply that a particular solution is more or less correct, but we can make the hypothesis that if larger positions are present in the data and the adopted similarity measure is appropriate, these will be identified with less need for approximations. This hypothesis is at the basis of our quantitative evaluation, discussed in Section 4.3. Fig. 5 depicts all the positions found in our working example on being part of the shortest path connecting pairs of nodes varying the relative cutting parameter h.



(a) h = 0% (b) h = 28% (c) h = 34% (d) h = 40%

(e) h = 41% (f) h = 46% (g) h = 49% (h) h = 52%

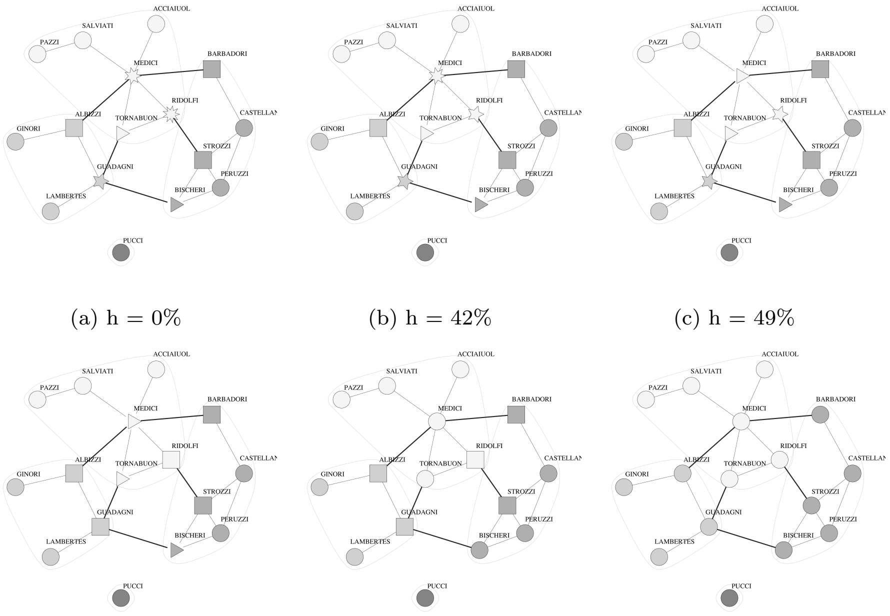
(i) h = 56% (j) h = 64% (k) h = 82% (l) h = 100%

Fig. 5: Positions detected as being part of the shortest path connecting pairs of nodes in the Marriage social network of Florentine families, at varying levels of approximation (color figure in the online version).

Fig. 5a represents the positional clustering with no approximation ($h_{min} = 0$), while Fig. 5l represents the other extreme where all nodes are placed in the same position. An approximation of 34% would be needed to change these initial groups.

A close observation of the clustering sequence in Fig. 5 highlights another interesting fact, which is also true in traditional blockmodeling: the positions found with no approximation can be misleading, even if they are formally correct. In Fig. 5a families *Acciaiuol*, *Ginori*, *Lambertes*, *Pucci*, *Pazzi* are detected in the same position because none of them is embedded in any shortest path between two other families. In practice, this is the position of nodes not captured by the comparison function, having only 0s in the corresponding row of the matrix.

Therefore it is advisable to consider this position as different from the others, or to directly remove empty rows from the extended similarity matrix M before computing the clustering to find positions or roles. By doing so, the approximation needed to find relevant positions with more than one node (Fig. 5c) would decrease from 34% to 17%.



(a) h = 0% (b) h = 42% (c) h = 49%

(d) h = 73% (e) h = 93% (f) h = 100%

Fig. 6: Positions detected as facilitating the exchange of information with the same communities in the Marriage social network of Florentine families, at varying levels of approximation.

Back to our example of positions detected as facilitating the exchange of information across the same communities, the sequence of h-cuts (Fig. 6) shows that, as we expected, with $h_{min} = 0$ we are immediately able to find meaningful positions.

3 Discovering new patterns in real social networks

As we have discussed early one, structural patterns might arise in (social) networks of any type and on different scales. The current blockmodeling methods have been demonstrated very effective to identify local patterns (e.g., structurally and regular equivalent actors), but they are not flexible enough to find structrural relations related with other groups of actors not directly connected to them (e.g., equivalent actors based on distance-based nodes' attributes).

In this section we apply our method to different social networks to perform a qualitative evaluation, and also to show how new patterns can be discovered by identifying the right equivalence measure on each particular case. Therefore, the similarity measures presented are just a sample of the new possibilities when we are able to compare actors with sets of actors. The identification of other positions and roles will probably require different measures. We divide our analysis in different parts, each one focusing on a different network.

First, we will use a synthetic network with a clear community structure to show the effect of including or excluding the own community set in the extended matrix calculation, e.g., including or excluding all the subsets $S'_k \supseteq \{i\}$ when the extended comparison function $D(i, S_k)$ is calculated. Then, we apply the method based on communities to detect affiliations in an academic social network with ground truth. Finally, we explore further the group formation of subsets showing how contextual information about the social network can be used to form more meaningful extended matrices.

3.1 Datasets

In the following we describe each of the networks used in our evaluation. Table 2 describes the main properties of the four networks.

The **Community benchmark** is a synthetic network built using the package for generating benchmark graphs with overlapping communities described in [Lancichinetti and Fortunato(2009)]. The parameters have been set to generate a network with five dense communities weakly connected to each other — only five nodes are connected to more than two communities.

Our **DBLP** network is a sub-graph extracted from the online computer science bibliography[2] using the XML API and manually verified. The DBLP sub-graph represents the co-authorship network of three of the authors of this work collected in December 2015. Therefore, all the nodes in the network are at maximum one hop from one of the authors, generating a network of diameter 6. The network also includes the collaborations between the 1-hop authors, but no other academics. We have checked that the network does not include any duplicated author, but we have not taken any action regarding possible confirmed-unconfirmed alias.

The **AUCS** [Rossi and Magnani(2015)] network contains a five-relational graph describing the social relations among employees of a Computer Science department. The five relations are: lunch, work, co-authorship, leisure and Facebook friendship. For the analysis we have created two distinct social networks using the work and co-authorship relations.

Table 2 describes the main properties of the four networks.

Table 2: Descriptive measures for the real social networks used: number of actors, Density, Clustering Coefficient, Degree Centralization, Average path length, Diameter

Network	**N**	**Dens**	**CC**	**DCentr**	**Avg Length**	**Dia**
Community benchmark	*300*	*0.016*	*0.05*	*0.007*	*5.66*	*10*
DBLP	*74*	*0.12*	*0.63*	*0.55*	*2.28*	*6*
AUCS Work	*60*	*0.11*	*0.34*	*0.35*	*2.39*	*4*
AUCS Coauthor	*25*	*0.07*	*0.43*	*0.14*	*1.50*	*3*

[2] http://dblp.uni-trier.de

The synthetic network used in this study offers an unequivocal perspective of community formation, with clearly identified positions to compare different extended measures. As we know the identity of the nodes interconnecting each of the communities, we can easily verify the results of our experimentation.

Both social datasets — DBLP and AUCS — describe, instead, collaborative social networks from two different perspectives. On one hand, DBLP describes collaborations among geographically distant people and therefore contains authors with affiliations to different research centers and universities. The two AUCS networks, on the other hand, have been collected among the members of the same research department and reflect its internal structure (research groups and academic positions).

3.2 Analyzing communities

Graph communities are, as we described above, cohesive groups of actors with many connections inside each group and fewer relationships with other actors outside them. In communication networks, for example, they might represent clusters of computers who are exchanging protocol information mainly between them. In a social context, instead, communities might represent users regularly talking about similar topics or hanging out together regularly.

Both examples describe real situations where the detection of communities and, most importantly, detecting their gateway members, is of interest. However, the analysis of both scenarios might seek different objectives: in the first case, we might be interested in sharing protocol information to specific clusters, while in the second case we could just be interested in detecting actors participating in similar activities without being interested in their original community. These are two examples showing different ways to describe positions based on equivalence relations between individuals and communities. While the first case focuses on actors facilitating the exchange of information with other communities, the latter also considers connectivity within the community to which the actors belong.

Identifying such positions requires, additionally to a measure of equivalence able to compute how two actors are structurally similar, some measure able to compare their similarity with(in) a subset of actors (here, a cluster of computers or a community of actors); which is not possible without changing the traditional blockmodeling framework.

3.2.1 Similarity measures only comparing relations with other communities

Fig. 7 shows the result of a positional analysis in the benchmark network. Each community is shaded with different background colors, and each vertex is colored according to its position at $h = 0\%$. Some relevant positions have been marked with labels, so that they can be discussed without relying on the colors.

The algorithm properly groups the five nodes in the center of the figure into three positions, based on their connectivity to and within their own community. As an example, one position (position P_1 in the figure) identifies the unique actors from the upper communities (C_1, C_2 and C_3) connected to the three bottom communities (C_4 and C_5). This position would not be identifiable without the extended similarity matrix, e.g., using the traditional concept of structural equivalence both

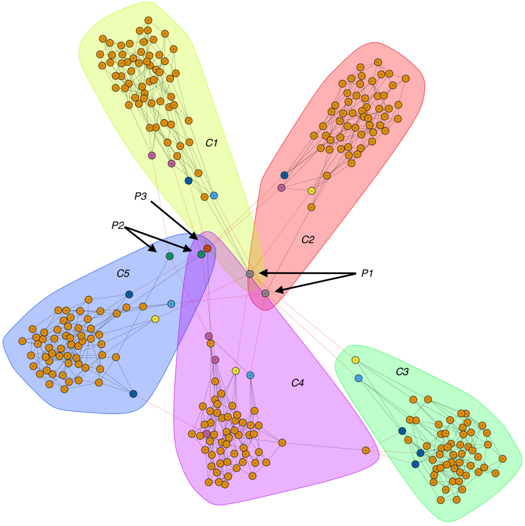


Fig. 7: Positions in the community benchmark network based on the relations with other communities (color figure in the online version).

actors would be placed into different positions, as they are in fact connected to very different sets of nodes. Instead, within our approach they are interchangeable as they act as a gateway between these three communities and their own (even if they are part of different original communities). The two central actors marked as position P_2 in the figure, instead, have been put together by the method because they connect community C_5 (blue) with C_1 (yellow) and C_4 (pink).

This highlights two important aspects of the measurement: firstly, two nodes i and j can be placed in the same position or not with respect to a third community C_k independently of the community to which they belong, except if they are connected together — in this case their position would also depend on their original communities. Secondly, actors connected only to their own community are all considered being in the same position by this measure — technically, because the corresponding rows in the extended matrix contain only 0s.

3.2.2 Similarity based on relations within other and also the same community

In order to place actors in the same position when they are from different communities, but acting as a gateway between them, we need to substitute Eq. 8 with a simpler version:

$$D(i, S_k) = \begin{cases} 1 \text{ if } \exists k \in S_k, \{i, k\} \in E \\ 0 \qquad\qquad\quad \text{otherwise} \end{cases} \tag{10}$$

In practice, when we compare each actor i with a subset S_k, now we are not checking if i is or not in the subset S_k, as it happens using Eq. 8. We simply

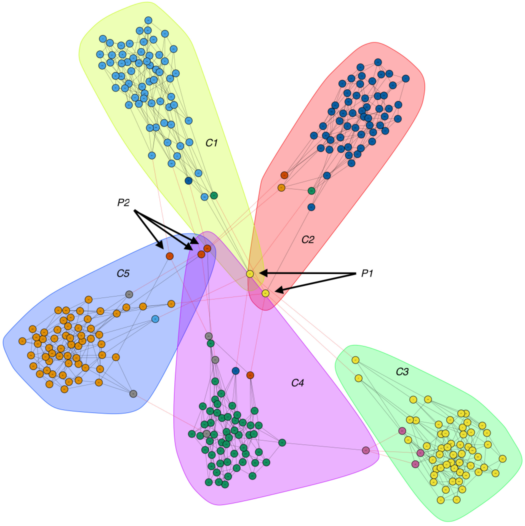


Fig. 8: Positions in the community benchmark network based on the relations within the same community (color figure in the online version).

check its inter-community connectivity. Fig. 8 allows us to compare the previous positional analysis with this new one.

We can observe that now nodes from different communities interconnected among them are considered in the same position like, as an example, the two nodes pointed by arrows. To better understand the difference consider three authors (namely, i, j and k) from three different communities (C_i, C_j and C_j). If i and j are both connected to k, according to the new measure both will be considered in a different position as i will have a value of 1 for C_i and C_k, while j will have 1s for C_j and C_k. Using the previous measure, instead, i and j would be considered in the same position as both would have just one 1 in their row, indicating their connectivity with C_k.

In the next section we explore the differences between these extended measures using real data.

3.3 Analyzing co-author networks

Co-author networks are formed by linking together authors — academics in our case — who have written and published together some work. In the case of the DBLP dataset, these works are computer science related papers resulting from a cooperative effort among the authors. Therefore, each vertex of our network represents an author, and the edges connecting them represent mid-term and long-term collaborations.

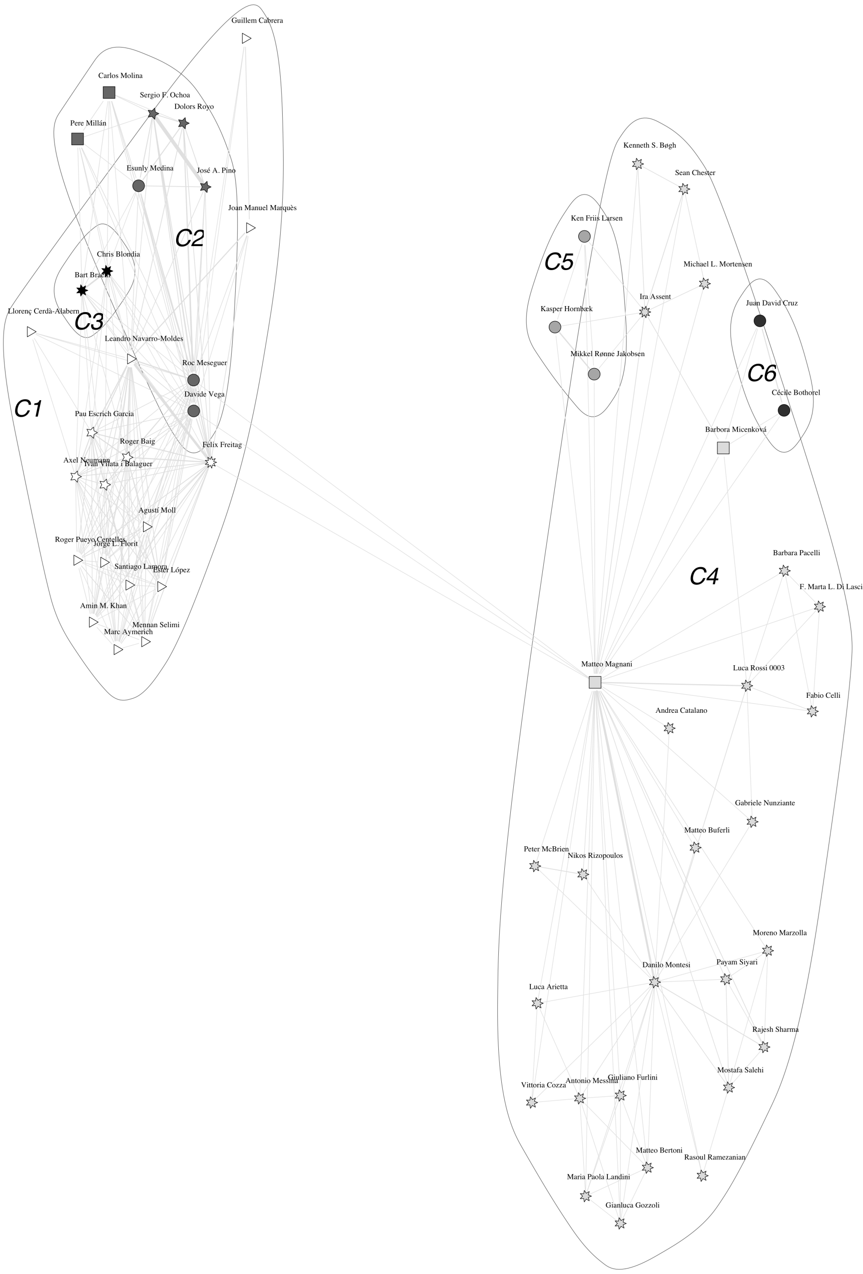


Fig. 9: DBLP Co-author ego-network of three authors. Each vertex represents an author, and the weight of the edges the number of publications in which both authors participated together. The highlighted areas correspond to the maximum modularity communities in the network, while the shape of the nodes indicates their positions with $h = 0$. Positions are detected using the binary community-based equivalence.

The network depicted in Fig. 9 represents a subset of the whole DBLP database, containing only the colleagues directly related with three of the authors of this work (treated as core-authors in the following). The network is divided into two main collaboration sub-networks — left and right side of Fig. 9 — describing the interactions between them at two different stages of their research careers. The members in the left side are mainly connected as participants of an European research project (CONFINE[3]), and include researchers from five different institutions. The right-side subnetwork, instead, describes a combination of long-term collaborations among researchers from Italian universities reinforced by the mobility of one of the authors across different research institutions.

Each of the authors in Fig. 9 is drawn inside its community obtained using a modularity maximization community detection algorithm (see Table 3). In the left-side, the larger community (C_1) contains most of the members of the research group hosting one of the core-authors who are working in a different campus. The second larger community (C_2) represents the members of the research group on the same campus and academics from other universities with whom he has regularly collaborated. Community C_3 identifies two academics from the University of Antwerp.

In the right-side sub-graph, the larger community (C_4) identifies researchers connected to the other two authors, who hold more senior academic positions than the other one and mainly includes researchers from universities located in Italy. The smaller communities in the right-side sub-graph represent two different research groups at universities where the author did not have a position, but only collaborations — in Denmark (C_6) and France (C_5).

These communities have been used as the subsets of actors to identify positions of interest using the already described detection of actors facilitating the exchange of information with the same communities, which are summarized in Table 3. We call the resulting positions *binary-community equivalence* to differentiate them from other possible measures based on grouping authors by communities.

Table 3 shows that the positional analysis performed using binary-community equivalence has been able to identify correctly all the authors' affiliations in the right-side sub-graph, thanks to the mobility pattern of their core-authors. Therefore, we can identify *Ira Assent* and *Barbora Micenkova* as two researchers providing collaboration opportunities between the authors from the Italian universities and the research centers from two different countries. The author with higher mobility is identified himself in a single position, as he acts as bridge between the two subnetworks.

Some of the affiliations detected in the left-side of the network, however, are incorrectly identified because the authors had only sporadic collaborations — recall that the main author in this subgraph is a junior researcher. As an example, the positional detection algorithm places *Guillem Cabrera* and *Joan Manel Marques* — which could be considered being in the *CONFINE project members* position — in the same position as the *local research group* despite being from another university; because the collaboration in the project from these two authors is limited to a very small subset of the project's members.

Similarly, one PhD student — who has the same advisor as this author — appears identified among the *members of Rovira i Virgili University*, because she

[3] https://confine-project.eu/

Table 3: Positions detected using the binary-community equivalence in a DBLP author-centric network.

Position	**Community**	**Authors**	**Errors**	**Description**
Triangles	C_1	*11*	*2 (18,2%)*	*Local research group. Barcelona*
5-stars	C_1	*6*	*0*	*CONFINE project*
Circle	C_1	*1*	*0*	*1st supervisor of core-author 1*
Circle	C_2	*2*	*0*	*Core-author 1 + 2nd supervisor*
5-stars	C_2	*3*	*1 (33,3%)*	*Rovira i Virgili University*
Triangles	C_2	*3*	*1 (33,3%)*	*University of Chile*
5-stars	C_3	*2*	*0*	*Antwerp University*
7-stars	C_4	*25*	*0*	*Author 2 + Italian collaboration network*
Square	C_4	*1*	*0*	*Author 3*
Sun	C_4	*1*	*0*	*Supervisor of author 3. Denmark*
Circle	C_4	*1*	*0*	*Co-author of author 3, w/ same supervisor*
7-star	C_5	*3*	*0*	*French collaboration network*
Circle	C_6	*2*	*0*	*Other Danish collaboration network*

collaborated with their same community. However, if we had included in the network formation also her complete network of coauthors — meaning, other members not directly connected to any author of the article — we would had detected her in a different single position, as she has written some works with people from other universities too. The error in the identification of her position is, then, caused by the lack of information about the network.

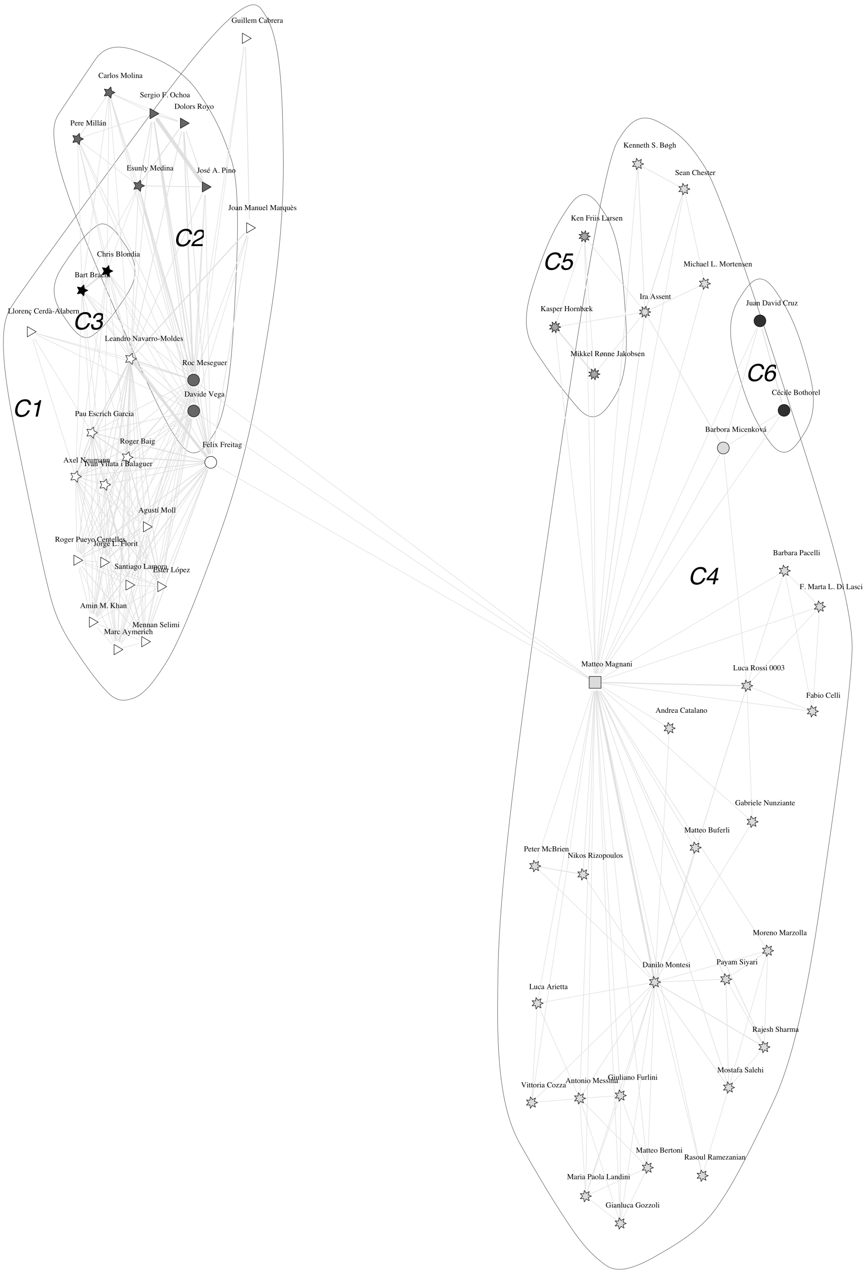


Fig. 10: DBLP Co-author network of three authors. Each vertex represents an author, and the weight of the edges the number of publications in which both authors participated together. The highlighted areas correspond to the maximum modularity communities in the network, while authors are colored according to their structurally equivalent positions with $h = 0$. Positions are detected using the binary community-based equivalence counting the own one.

Given the missing information, one might consider including the actors' own community in the extended similarity measure, using Eq. 10, as we did with the community benchmark.

Compared with the previous analysis, we can see now that positions are determined by the communities to which authors are collaborating too, grouping together previously disjoint communities. As an example, *Felix Freitag*, one of the supervisors of the author in community C_1 who was previously placed in a different position (because his working place is in a different campus), now is detected in the same position as the author. Similarly, authors from different affiliations (like, the academics from *Antwerp University* and *Rovira i Virgili University*) are now detected in the same position as other *CONFINE project active members*.

This second analysis no longer considers the affiliation information, but provides more significant positions based on the collaboration between authors.

3.4 Analyzing co-workers networks

So far, in this section we have built extended similarity matrices by selecting the subset of actors S using some structural properties of the graph, meaning the communities with maximum modularity. Despite the good results obtained in the previous cases, this method can lead to inconclusive results for some particular networks. Fig. 11 represents the co-authorship layer of the AUCS network. Each vertex represents a single member of the same computer science department, and an edge between them represents a research work published together. For simplification we have deleted all the members with no ties to other colleagues.

Compared with previous networks, we can observe that the AUCS co-authorship graph has fewer vertices, grouped into 8 different components with six of fewer vertices each. Therefore, a community detection algorithm will probably match each network component with one community, with no inter-communities ties. The extended similarity measures will, then, identify all nodes in the same — non interesting — position.

A more suitable analysis, however, would be to use context information provided by the AUCS network, and perform a more meaningful analysis taking advantage of the flexibility of our framework. Specifically, the AUCS dataset provides the research group/s and the academic rank for every actor in the network. So, we intended to use the "research group" information to generate the subsets S of actors, and then check positions as actors facilitating the exchange of information with the same communities (which in this context, will be the research groups). This is possible, because despite being related, the extended similarity measure D and the sub-setting S are two different components of the framework.

Fig. 11 shows the results of this experiment, grouping nodes within the same research group together and identifying their position by shape. Nodes' labels identify also the ground truth positions provided by the dataset. We can observe that our framework identified as single-actor positions almost all the professors in the research group and all the administrators involved.

From a context-aware point of view, the result seems logical. Professors (or higher-ranked academics) and administrators are both academic positions which will be more frequently involved on projects between several research groups. The

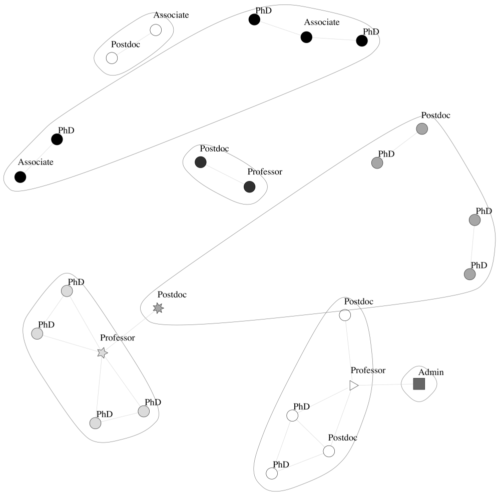


Fig. 11: Co-author network with research groups (areas), academic positions (names) and detected positions (shapes)

former ones, as leaders of the research; while the last ones as administration coordinators. In order to further test this idea, we ran the same experiment using the AUCS work graph that includes all the researchers collaborating in some common project (and not just writing some paper together); and we will detect *positions* and *roles* of being or not a higher-ranked academic or administrator. That is, we cut both clustering dendrograms into two single groups based on the similarity measure.

3.4.1 Comparing higher-ranked academics and administrators.

Fig. 12 compares the different positions and roles detected using this method with the betweenness centrality of the actors. Instead of showing the resulting networking or blockmodels, we have grouped the different classes into a single heatmap to compare the clusters more easily. Each row of the heatmap represents, then, a single actor in the AUCS network, while the three main columns represent the positions, roles and the betweenness centrality of such actors. Additionally, in the left-most side of the heatmap we added the truth role of each actor according the literature.

According to this result, the positional analysis identifies correctly 95.02% of the professors, associate professors and administrative personnel in the department under the correct position, and incorrectly one very proactive PhD Candidate. We also believe, given the lack of information, that the actor with *unknown* ground-truth position has been correctly identified among higher-ranked academics. Comparing the results of the positional analysis with the role assignment, the second one makes an additional mistake incorrectly placing a PostDoc researcher.

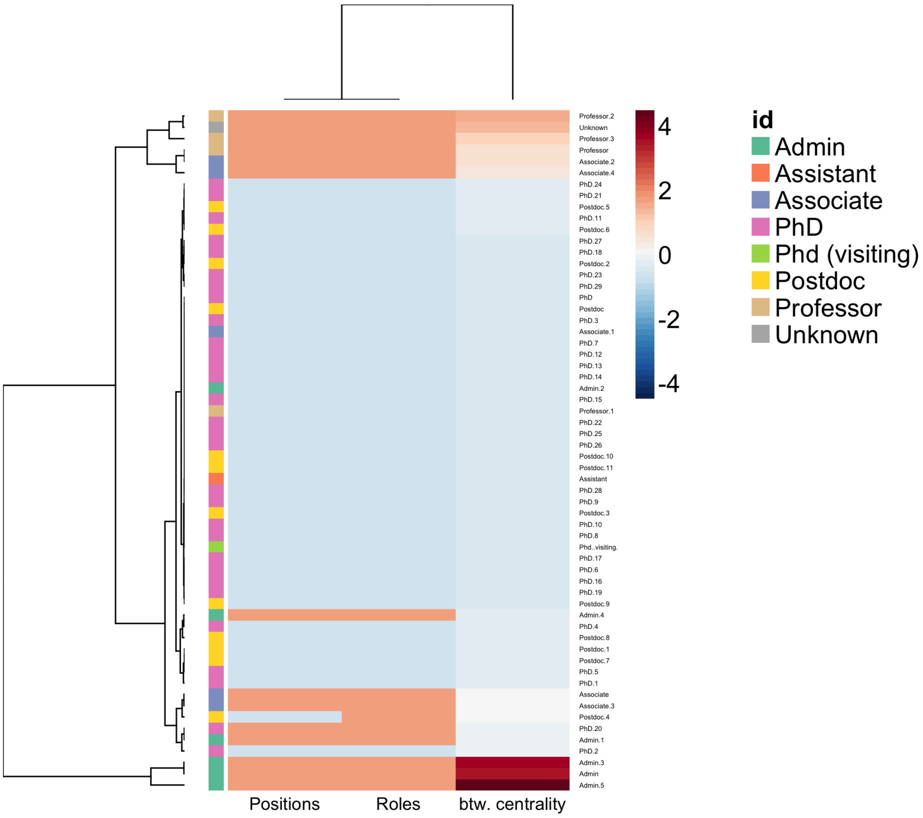


Fig. 12: Co-workers: comparison between positions, roles and betweenness centrality.

We believe that compared with common network analysis measures, position and role analysis provides further insights about the structure of the network, as it is able to connect structural patterns with extended measures. For example, while is true that an analysis of the AUCS work graph based on centrality measures (See Fig. 12) would identify some of the members of the positions found using our approach, it is also true that it would ignore 33,33% of the actors which we would be interested in.

4 Framework validation

The evaluation of new structural patterns is usually conditional on demonstrating that they are actually found. However, as the optimization function we have proposed is the same as the evaluation function, this would be always certain;

and therefore worthless. Instead, the approach we have chosen in this section is to compare the resulting patterns found by our framework using two representative equivalence measures in real social and synthetic networks, aiming to demonstrate that the positions found do not emerge by chance and, hence, provide a significant description of distintive interaction patterns found in the analyzed datasets.

Therefore, the objective of this section is twofold: firstly, we want to measure qualitatively how each of the components of our framework impacts the outcome of positional detection; and in second term, we want to discuss how the different extended similarity measures are affected by the network in which they are applied. We got an intuitive idea about both questions in the previous section — when we analyzed the AUCS co-workers network. However, in order to make the following results as generalizable as possible we will constrain our analysis to synthetic networks with well-known properties; and therefore use real social networks only for comparison purposes.

To evaluate our framework we built a library in R using the *blockmodeling-package* [Žiberna(2007)] and the *fastcluster library* [Müllner(2013)] as baselines. The library has been used to perform the experimental analysis presented in this work, which included the detection of roles and positions with different combinations of networks — real or synthetic — and similarity measures — traditional or extended. The library also provides functions to plot and analyze sequences of positions with different degree of approximation as we will see in Section 4.2.

4.1 Datasets

We evaluated quantitatively our proposal using a set of synthetic networks, based on the **Erdos-Renyi** model [Erdős and Rényi(1959)] (ER) — also known as $ER(n, p)$. The synthetic graphs used are different in density and connectivity, but not in size; allowing us to easily compare the behaviour of different similarity measures on multiple scenarios. For an Erdos-Renyi model we can predict the formation of a giant component and the average degree of the network. Recall that in ER models the giant component starts appearing when the probability p of two nodes being connected is higher than the threshold n^{-1}. In the same way, after the threshold $p \geq \frac{ln(n)}{n}$ the network is likely to be completely connected.

We have also used a flattened version of the **AUCS** [Rossi and Magnani(2015)] network, which contains in the same mono-relational graph all edges present on any of its 5 original layers — which will match the number of selected vertices n in the ER models.

Table 4 describes the main properties of the real social network, and the average main properties of the 10 different Erdos-Renyi networks used for each probability.

Again, the size of the synthetic graphs used in the framework evaluation is subordinated to the size of our baseline (or real) network, which describes a social environment already described in the literature; which will make the evaluation of the results easier.

Table 4: Descriptive measures for the networks used: number of actors, Density, Clustering Coefficient, Degree Centralization, Average path length, Diameter

Network	**N**	**Dens**	**CC**	**DCentr**	**Avg Length**	**Dia**
AUCS Flatten	*61*	*0.034*	*0.048*	*0.048*	*2.06*	*4*
ER(p = 0.01)	*61*	*0.010*	*0.07*	*0.037*	*1.77*	*4.13*
ER(p = 0.02)	*61*	*0.020*	*0.01*	*0.051*	*3.91*	*9.49*
ER(p = 0.03)	*61*	*0.029*	*0.03*	*0.059*	*4.98*	*12.01*
ER(p = 0.05)	*61*	*0.051*	*0.05*	*0.072*	*3.62*	*8.03*
ER(p = 0.07)	*61*	*0.069*	*0.07*	*0.087*	*2.95*	*6.23*
ER(p = 0.09)	*61*	*0.090*	*0.09*	*0.092*	*2.57*	*5*

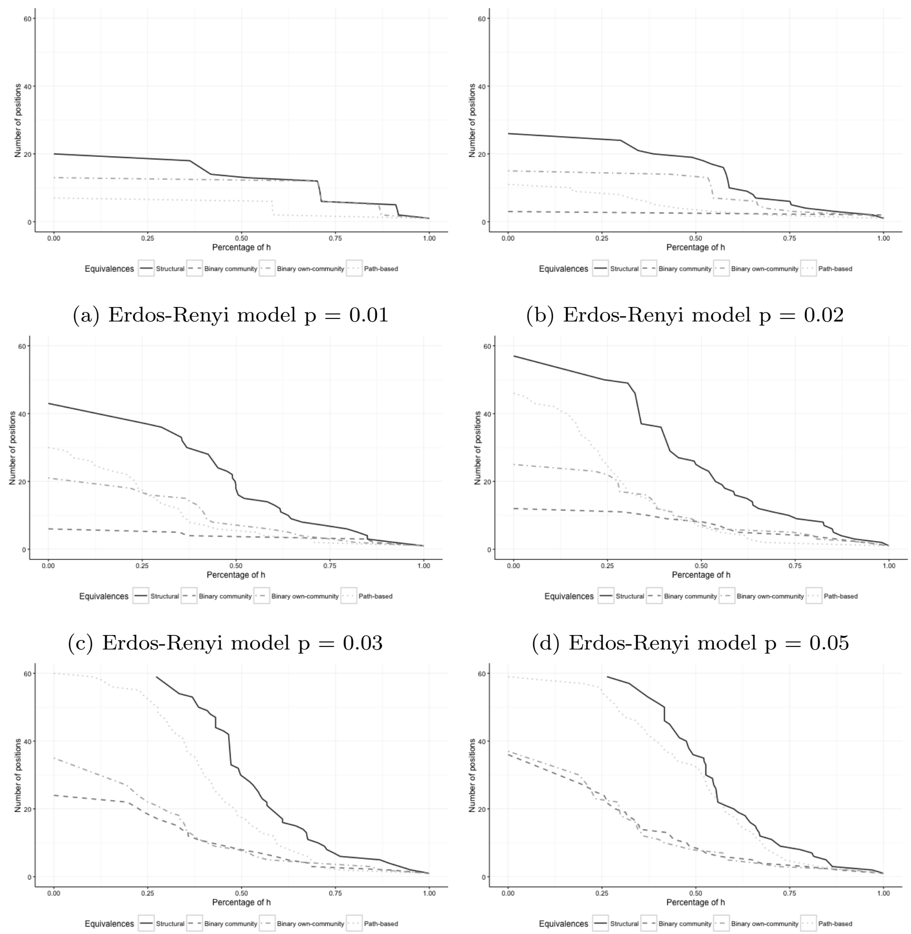
4.2 Similarity measures analysis

In this work we introduced a framework to find new roles and positions associated to group relations in a social network. As any indirect blockmodeling methodology, the meaning and profitability of the findings are highly tied with the similarity measure used, which has to be chosen carefully. Potentially, any measure based on the topological structure of the social network — like distance-based measures — could be used as an equivalence. However, in practice, it is necessary to discard measures of equivalence that do not find any dissimilarity — meaning, all actors are always grouped together — or measures where the assignment of actors into either roles or positions is a consequence of some random phenomenon. We have observed this phenomena during our study of the AUCS co-author and co-workers networks in Section 3.

While the first problem is more dependent on the network structure (e.g. by definition we cannot find more than one position or role in Torus networks), the later is mostly related with our framework and hence we need to address it. In order to verify the lack of randomness in our equivalences we tested the framework by comparing the positions found in six ER different random graphs. Each synthetic graph of our testbed contains 61 vertices and a different probability p between 0.01 and 0.09.

The experiment consisted on computing, for each of the mentioned graphs, all possible positions using the structural equivalence and the set of extended equivalences presented through this work. The **path-based** positions, therefore are finding *actors being in exactly the same shortest path that connects pairs of nodes*, while the **community-based** positions are detecting *actors facilitating the exchange of information with the same communities* counting their own community — **community-own based** — or not. Each experiment has been repeated 10 times in order to avoid random effects caused by the network formation.

Fig. 13 shows, for each percentage of connectivity p, the number of positions found as y-axis and the corresponding approximation level h (x-axis). Hence, the left-most value of h for each curve represents the minimum approximation — maximum exactitude — found in the measure (h_{min}). We can observe that in random networks with not all vertices connected to the same component — which have $p < 0.067$ — all analyses find exact positions ($h_{min} = 0$). This is reasonable,



(a) Erdos-Renyi model p = 0.01

(b) Erdos-Renyi model p = 0.02

(c) Erdos-Renyi model p = 0.03

(d) Erdos-Renyi model p = 0.05

(e) Erdos-Renyi model p = 0.07

(f) Erdos-Renyi model p = 0.09

Fig. 13: Number of different positions found as the percentage of approximation for three extended measures and equivalence. Each plot represents a different synthetic graph based on the Erdos-Renyi model with 61 vertices and different connectivity p.

as actors are grouped together based on their connectivity with other elements of the network (actors, paths or communities). Instead, in the $ER(61, 0.07)$ and $ER(61, 0.09)$ models, the structural equivalence measure cannot find "exact" positions.

The different sizes of positions depicted in Fig. 13 are related with the flexibility, in terms of measure, of each equivalence. More sub-settings of actors will increase the number of possible row combinations, and hence lead more heterogeneous positions; while fewer sub-settings will tend to create less positions with a smaller approximation. Another factor that may influence the flexibility of the measure is how the connectivity between actors and subsets is defined. As an example, for a given actor it is different to count if it is connected or not to a subset of actors (e.g., a community) or to count if it is embedded in one of the shortest paths that two actors are using to communicate to each other. The measures proposed in this work are good examples of meaningful measures for the analysis of social networks, but the framework does not constrain the definition to specific measures or sub-settings.

Another important aspect that must be taken into account for the interpretation of the results is that the indirect blockmodeling methodology needs some care in order to distinguish between positions — or roles — related with the similarity measure, and other clusters containing actors not really captured. Consider, for example, our path-based extended measure, which tries to group together actors in the same shortest path as other actors. Then, by definition, nodes that are not in any shortest path or are completely disconnected to other nodes in the graph will be all placed in the same position by the clustering algorithm. However, they do not represent a position of interest, but rather a set of actors for which the measure does not apply.

4.3 Evaluation of the approximation

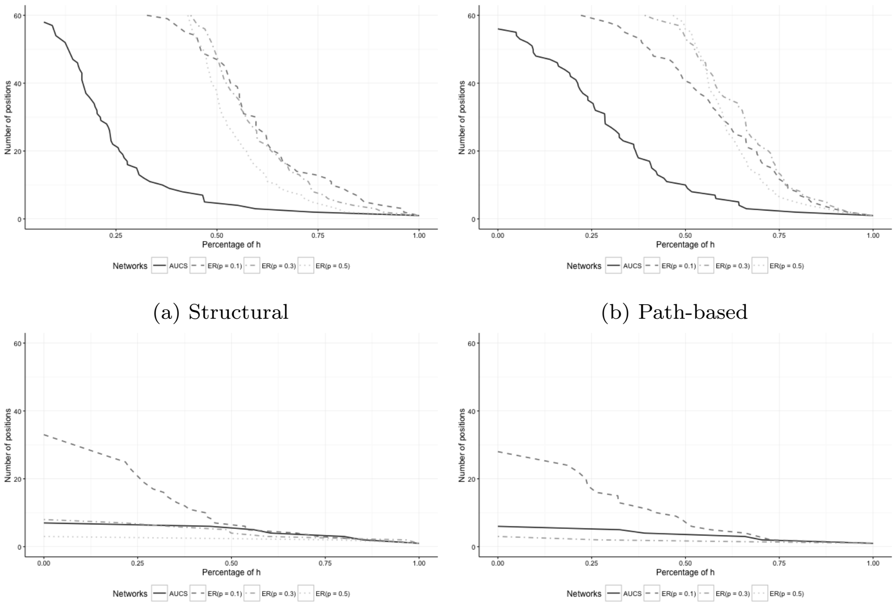
Our previous analysis has focused on the different approximation levels generated by our framework once applied to some ER random graphs. It is known, however, that synthetic models — like the ER — are intended for limited purposes and generally only capture a fraction of the inherent characteristics of real social networks, like degree distribution, clustering coefficient and clique distribution.

Therefore, by comparing the different positions found in synthetic ER graphs with the positions found in some real networks like AUCS, we expect to unveil hidden properties that will make the real networks' positions more "meaningful". Fig. 14 shows the results after repeating the last experiment in several synthetic graphs with higher clustering — observe that all of them are above the threshold for a single component — and a flattened version of the AUCS real network.

We can observe that, in general, the *structural* and *path-based* first positions detected — those considered more meaningful by the clustering algorithm — in the real AUCS network are more than 30% less approximate than any Erdos Renyi graph.

It is our belief that the fact that positions found in real social networks (like AUCS) have smaller h_{min} indices, is an indication that these positions are intrinsically more embedded in the structure of the network, and it is worth to study them.

More complex to understand are the results referring to the community-based positions in Figures 14c and 14d, where the indices of approximation for highly connected ER models do not highlight any concrete pattern — which we have confirmed by manually checking the results on each network. The main reason behind it is the expected lack of community structures in ER model networks, particularly in networks with fewer ties (e.g., the modularity measure of all ER networks, except the $ER(61, 0.1)$, are below 0.02). As a consequence, the similarity measure is not able to capture patterns because it is actually comparing relationships between single actors and groups of actors who are not embedded together



(a) Structural

(b) Path-based

(c) Binary community-based

(d) Binary own-community based

Fig. 14: Number of different positions found as the percentage of approximation for three extended measures and equivalence. Each line represents the AUCS network and different synthetic graphs based on the Erdos-Renyi model with 61 vertices and different connectivity p.

in a community. The $ER(61, 0.1)$ model networks, instead, have significant modularity values (larger than 0.7); and hence is expected that the patterns of relations found are more significant.

5 Related work

5.1 Blockmodeling

Both concepts of role and position have been redefined many times in the literature, by mathematicians and sociologists, more or less formally. In our work we use the definition provided by Wasserman and Faust [Wasserman(1994)]:

> "*In social network analysis* ***position*** *refers to a collection of individuals who are similarly embedded in networks of relations, while* ***role*** *refers to the pattern of relations which obtain between actors or between positions. The notion of position thus refers to a collection of actors who are similar in social activity, ties, or interactions, with respect to actors in other positions.*"

Despite the lack of mathematical notation, this definition clearly states the idea of identifying positions as a clustering problem where actors — vertices of a graph — are assigned to smaller subsets — called positions — based on a notion of similarity. The similarity not only measures how similar the local connectivity between pairs of actors is in the graph, but can also measure other properties or relations.

Structural equivalence is the most basic and strict notion of similarity. Other similarities have been later developed to relax the notion of equivalence. In *regular equivalence* [Wasserman(1994)], for example, two actors are in the same position if they have similar relations with other positions; while in *automorphic equivalence* [Borgatti(1992)] two nodes (i, j) are equivalent if there is an automorphism ρ of G where $i = \rho(j)$. To mention a third well-known example, in *stochastic equivalence* [Doreian et al(2005)Doreian, Batagelj, and Ferligoj] two actors are in the same position if they have the same probability distribution of ties with other actors, which is more similar to the notion of role that we are presenting in this work.

In a general form, we can generalize these notions of *equivalence* as a set of *node-based* features: $A_h(i) \quad \forall h \in H$ and *distance-based* features: $D_p(i,j) \quad \forall p \in P; k \neq i, j$; where H is a set of node attributes and P is a set of comparison functions. Notice that these relations still constrain the model to a) the adjacency connectivity matrix and b) pairwise actor comparisons (i, j). For an extended taxonomy and classification we suggest the recent work of Rossi and Ahmed [Rossi and Ahmed(2015)].

Several generalizations have been developed to find positions without perfect similarity/dissimilarity [Doreian et al(2005)Doreian, Batagelj, and Ferligoj], to be used in weighted graphs [Žiberna(2007)], or even to find non-trivial equivalent positions [Brusco et al(2013)Brusco, Doreian, Steinley, and Satornino]. While these approaches have proved useful to detect some kinds of positions, and are flexible enough to accommodate different kinds of similarity functions, they are also based on pairwise relationships. However, the general idea of finding approximate equivalences is also fundamental in our framework, because a strict check for equivalence would rarely identify any groups of similar actors in real social networks.

5.2 Discovery and mining

Apart from the literature related with blockmodeling, there is a large number of works related with role detection in networks focused on *role discovery* and *graph mining*. Like direct blockmodeling — and unlike our proposal — these techniques do not assume any previous knowledge about the network under analysis. However, they differentiate themselves from direct blockmodeling because they use machine learning techniques to find actors with similar patterns of connectivity — meaning, roles.

One of the earliest works worth mentioning in this area is SimRank, a scoring algorithm to identify regular equivalent roles under some notion of *context similarity* [Jeh and Widom(2002)]. This similarity measure is computed recursively according to the average similarity of all the neighbour pairs. That is, two actors will have the same role if they are connected — on average — to the same number of nodes from each other context (an actor feature). The method was improved lately by Jin et.al. [Jin et al(2011)Jin, Lee, and Hong] to guarantee that two nodes will be considered in the same role if, and only if, they are *automorphically equivalent*. Both proposals were aimed to discover roles in larger graph structures than blockmodeling techniques, but they limit their discoveries to regular and automorphic roles.

More recently, RolX [Henderson et al(2012)] was proposed in order to generalize the roles detected across networks. In order to maintain its scalability when the number of features increases, RolX decomposes the adjacency matrix of the original graph into two matrices (node-role and role-feature) using non-negative matrix factorization (NMF) for inferring the roles from the set of features. Then, it explores the new matrices using transfer learning. A similar technique was used in GLRP [Gilpin et al(2013)Gilpin, Eliassi-Rad, and Davidson] to discover roles with supervision. While both proposals are able to find roles with similar structures in larger graphs than our proposal — and generally, any other blockmodeling system — they cannot relate this structures with the local structure of the network without losing precision.

Panther [Zhang et al(2015)Zhang, Tang, Ma, Tong, Jing, and Li] focuses on increasing the performance and space memory of the previous algorithms. Specifically, instead of computing the structure similarity of each feature in the graph, Panther performs a fixed set of random walks (R) over the network starting from a randomly picked vertex and walking another fixed number of steps (T). If both, R and T, are sufficiently large, Panther guarantees that the sampling result will be accurate enough to represent the structure of the network. Then, the algorithm computes a score for each pair of nodes on each random walk. Even if their approach is conceptually different from ours, with minimum changes the algorithm would be able to find roles with similar meaning as our roles based on *being in the shortest path* just by counting the number of random walks where each actor participate.

In [Rossi and Ahmed(2015)] the authors describe a new taxonomy for role discovery methods, which also introduces the idea of "feature-based role discovery" as a subclass of the previous techniques presented. An updated and complete survey about discovery and mining methods and classification can be found in [Rossi and Ahmed(2015)]. According to these proposals, the similarity between nodes can be measured using a set of node-structural features (e.g. degree, distance, etc.), which can be any set of measures taken from the initial graph. Together, they create a new matrix — or a set of orthogonal matrices — containing all the measures related to the actors and their features. Then, they use machine learning techniques to infer the social feature-based roles.

It is possible to argue that some of the extended equivalences proposed in our work could be used as features in these models, but in our framework we keep track of the relation between the measure — or feature — and the nodes related to it — the subsets of nodes that are needed to compute the measure. Because of this, our framework is able to measure not only patterns of relations (roles), but also positions. More importantly, we are able to relate both concepts to the same measure. However, *feature-based role discovery* techniques have been demonstrated more efficient with respect to time and computational resources for larger networks.

6 Conclusion and discussion

Blockmodeling has been primarily used as a way to detect roles and/or positions in social structures using node-based measures. Several extensions have been proposed for blockmodeling, many of which based on replacing the original comparison function with alternative ways of measuring the network structure around the nodes. Motivated by the will of applying this traditional approach to more complex network models that have recently received a renewed and extensive attention — like multi-relational networks — we have proposed a conceptual extension of blockmodeling that allows us to plug in additional comparison functions not usable in a standard setting.

To enable the usage of the additional types of similarity functions discussed in this work it is necessary to extend the regular similarity matrix into a more complex structure able to relate actors in a network with a) the extended measure and b) the extra information used to compute such measure — in this case, subsets of actors. These new measures generate a new asymmetric equivalence matrix that can be analyzed to find both social roles and positions.

In addition to a thorough presentation of our framework, exemplified and tested on several different types of networks, in this work we have focused on two kinds of measures, based on belonging to similar shortest paths between actors and being connected to similar groups. Through a qualitative analysis we have discussed how and why to use each of the measures, showing how the method can find meaningful positions in real networks. Additionally, we have shown that both measures can be enriched using context information about the social network (e.g., we used the research group information in the AUCS network to build the reference subsets of actors used to define positions).

6.1 Limitations

Compared with methods for role discovery and graph mining, our extended blockmodeling framework — and generally most of the actual solutions based on blockmodeling — presents a higher flexibility. In our particular case, this flexibility comes from the extension of the similarity matrices; and the many more measures they allow. However, the selection of the extended measures depends entirely on the objective of the analysis and/or the meaning of the desired positions and roles. Theoretically, any measure computed in a graph comparing a node to a set of nodes would be a candidate. In practice, we have observed that some of the measures require higher values of approximation (h) to identify positions containing multiple nodes, and of course the analyst should be aware of the semantics of the comparison function to be able to interpret the corresponding roles and positions.

Additionally, compared with other role discovery methods our method is aimed for smaller networks, where the practical size depends on the product between the number of actors and the number of groups.

6.2 Implications and future directions

Our clustering procedure is similar to the generalized blockmodeling for two-mode networks described in [Doreian et al(2004)Doreian, Batagelj, and Ferligoj], where each member of one mode is compared against the members of the other to compute a positional similarity. However, our extended similarity matrix differs from the two-mode adjacency matrix because each actor might be potentially present in one row and multiple columns. And hence, the simplified network resulting from the positional analysis must be represented as an hypergraph: each of the positions ρ will be present as a vertex of the new graph as well as all the original subsets S; which will be connected simultaneously to multiple positions. Understanding this new relation between roles and groups of original actors will require to develop new analysis tools.

Additionally, following our motivations, we want to extend the current proposal to other graph models, in order to understand more complex systems like multi-relational social networks. This next step will require the consideration of multi-relational measures.

Acknowledgment

This work has been partly supported by the Italian Ministry of Education, Universities and Research FIRB grant RBFR107725 and by the Spanish government under contract TIN2013-47245-C2-1-R.

References

[Borgatti(1992)] Borgatti SP (1992) Notions of position in social network analysis. Sociological methodology 22(1):1–35, DOI http://dx.doi.org/10.2307/270991

[Borgatti and Everett(1993)] Borgatti SP, Everett MG (1993) Two algorithms for computing regular equivalence. Social Networks 15(4):361–376, DOI http://dx.doi.org/10.1016/0378-8733(93)90012-A

[Breiger and Pattison(1986)] Breiger RL, Pattison PE (1986) Cumulated social roles: The duality of persons and their algebras. Social Networks 8(3):215–256, DOI http://dx.doi.org/10.1016/0378-8733(86)90006-7

[Brusco et al(2013)Brusco, Doreian, Steinley, and Satornino] Brusco M, Doreian P, Steinley D, Satornino C (2013) Multiobjective blockmodeling for social network analysis. Psychometrika 78(3):498–525, DOI "http://dx.doi.org/10.1007/s11336-012-9313-1"

[Coscia et al(2011)Coscia, Giannotti, and Pedreschi] Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining 4(5):512–546, DOI 10.1002/sam.10133

[Doreian(1988)] Doreian P (1988) Equivalence in a social network. Journal of mathematical sociology 13(3):243–281, DOI http://dx.doi.org/10.1080/0022250X.1988.9990034

[Doreian et al(2004)Doreian, Batagelj, and Ferligoj] Doreian P, Batagelj V, Ferligoj A (2004) Generalized blockmodeling of two-mode network data. Social Networks 26(1):29 – 53, DOI http://dx.doi.org/10.1016/j.socnet.2004.01.002

[Doreian et al(2005)Doreian, Batagelj, and Ferligoj] Doreian P, Batagelj V, Ferligoj A (2005) Generalized Blockmodeling. Structural Analysis in the Social Sciences, Cambridge University Press

[Durbin(1973)] Durbin J (1973) Distribution Theory for Tests Based on the Sample Distribution Function. Society for Industrial and Applied Mathematics, DOI http://dx.doi.org/10.1137/1.9781611970586

[Erdős and Rényi(1959)] Erdős P, Rényi A (1959) On random graphs. Publicationes Mathematicae Debrecen 6:290–297, DOI http://dx.doi.org/10.1234/12345678

[Fortunato(2010)] Fortunato S (2010) Community detection in graphs. Physics Reports 486(3-5):75–174, DOI http://dx.doi.org/10.1016/j.physrep.2009.11.002

[Gilpin et al(2013)Gilpin, Eliassi-Rad, and Davidson] Gilpin S, Eliassi-Rad T, Davidson I (2013) Guided learning for role discovery (glrd): Framework, algorithms, and applications. In: Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, pp 113–121, DOI http://doi.acm.org/10.1145/2487575.2487620

[Jeh and Widom(2002)] Jeh G, Widom J (2002) Simrank: A measure of structural-context similarity. In: Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '02, pp 538–543, DOI http://doi.acm.org/10.1145/775047.775126

[Jin et al(2011)Jin, Lee, and Hong] Jin R, Lee VE, Hong H (2011) Axiomatic ranking of network role similarity. In: Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '11, pp 922–930, DOI http://doi.acm.org/10.1145/2020408.2020561

[Lancichinetti and Fortunato(2009)] Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physics Reports 80:016,118, DOI http://dx.doi.org/10.1103/PhysRevE.80.016118

[Murtagh and Legendre(2014)] Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? Journal of Classification 31(3):274–295, DOI http://dx.doi.org/10.1007/s00357-014-9161-z

[Müllner(2013)] Müllner D (2013) fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. Journal of Statistical Software 53(1):1–18, DOI http://dx.doi.org/10.18637/jss.v053.i09

[Rossi and Magnani(2015)] Rossi L, Magnani M (2015) Towards effective visual analytics on multiplex and multilayer networks. Chaos, Solitons and Fractals 72:68–76, DOI http://dx.doi.org/10.1016/j.chaos.2014.12.022

[Rossi and Ahmed(2015)] Rossi R, Ahmed N (2015) Role discovery in networks. Knowledge and Data Engineering, IEEE Transactions on 27(4):1112–1131, DOI http://dx.doi.org/10.1109/TKDE.2014.2349913

[Vega et al(2015)Vega, Magnani, Meseguer, and Freitag] Vega D, Magnani M, Meseguer R, Freitag F (2015) Role and position detection in networks: reloaded. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[Wasserman(1994)] Wasserman S (1994) Social network analysis: Methods and applications, vol 8. Cambridge university press

[White and Reitz(1983)] White DR, Reitz KP (1983) Graph and semigroup homomorphisms on networks of relations. Social Networks 5(2):193 – 234, DOI http://dx.doi.org/10.1016/0378-8733(83)90025-4

[Zhang et al(2015)Zhang, Tang, Ma, Tong, Jing, and Li] Zhang J, Tang J, Ma C, Tong H, Jing Y, Li J (2015) Panther: Fast top-k similarity search on large networks. In: Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '15, pp 1445–1454, DOI http://doi.acm.org/10.1145/2783258.2783267

[Žiberna(2007)] Žiberna A (2007) Generalized blockmodeling of valued networks. Social networks 29(1):105–126, DOI http://dx.doi.org/10.1016/j.socnet.2006.04.002