Group Definition Based on Flow in Community Detection

María Barroso[1(✉)], Inmaculada Gutiérrez[1], Daniel Gómez[1,2], Javier Castro[1,2], and Rosa Espínola[1,2]

[1] Faculty of Statistics, Complutense University, Avenida Puerta de Hierro, s/n, 28040 Madrid, Spain
{mbarro10,inmaguti}@ucm.es, {dagomez,jcastroc,rosaev}@estad.ucm.es
[2] Instituto de Evaluación Sanitaria, Complutense University, Madrid, Spain

Abstract. Community detection problems are one of the hottest disciplines in social network analysis. Nevertheless, most of the related algorithms are specific for non-directed networks, or are based on a density concept of group. In this paper, we deal with a new concept of community for directed networks that is based on the classical flow concept. A community is strong and cohesive if their members can communicate among them. With the aim of dealing with the identification of this new class of groups, in this work, we propose the use of fuzzy measures to represent the flow capacity of a group. We also provide a competitive community detection algorithm that focus on the identification of these new class of flow-based community.

Keywords: Directed networks · Flow · Fuzzy measures · Community detection problem · Louvain Algorithm

1 Introduction

Community detection problems are one of the most important topic in social network analysis [7,18]. The idea of finding communities is strongly related to the idea of finding clusters in data analysis. In general, a cluster can be considered as a set of items that are *closer each other* when they are compared to the rest of items of the problem. A good clusterization of a set of items is associated with the identification of a set of clusters that present internally high degree of intra-homogeneity, and high degree of inter-heterogeneity. In networks, the idea of intra-homogeneity of a cluster/community is usually associated with the density of the group. Then, a good community will be a dense set of nodes with many connections between the members of each group. The idea of high inter-heterogeneity is associated with the existence of lower relations between the clusters/groups. So, the more relationships among the groups, the greater the

This research has been partially supported by the Government of Spain, Grant Plan Nacional de I+D+i, MTM2015-70550-P, PGC2018096509-B-I00 and TIN2015-66471-P.

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 524–538, 2020.
https://doi.org/10.1007/978-3-030-50153-2_39

degree of inter-heterogeneity is. Taking this into account, an optimal community may be a set of nodes that induces a completed subgraph (so they are strongly connected), and are isolated from the rest of the nodes of the network (so they have no relation with any other node in the network).

However, this idea of community is not unanimously accepted when the graph is directed and valued. As noted in [15,18], the idea of community in directed networks can have different interpretations, so several definitions could be made. This is the reason why finding clusters in directed networks is a challenging task with several important applications, since many of the real networks are modelled in an undirected way.

Despite the importance of community detection problems in directed networks, this problem has been poorly studied in the literature. Nevertheless, we can find in [18] four different (but non-formalized) concepts about what could be understood as a community/group:

- The first notion is about a random walk. In this case, each group is formed by these nodes that are more likely to remain inside than outside. This kind of communities are usually obtained with random walk techniques. In the literature, we can find some algorithms that deal with these problems (see for example [7,19,24]).
- The second notion is about the density. In this context, the groups of nodes follow the traditional clustering definition, based on edge density characteristics. It is important to mention that, in this sense, the concept of modularity [20] has been redefined for directed networks. Taking this into account, new algorithms have been developed to deal with this problem. Many of them are adaptions of some well-known community detection algorithms (as the directed Louvain [2]) to the directed case.
- Co-citation groups. As mentioned in [18], edges density is not always the only criterion to identify a set of nodes that share many characteristics. In directed networks, the idea of co-citation group tries to identify groups of nodes (not necessarily connected) that follow or are followed by the same groups. In this sense, we could have a group of nodes that form a community because its followers' set is the same, even if they do not know each other. We also could have a group of nodes that form a community because the set of nodes from which they extract the information is the same. If we think of a citation network, for example, a community could be formed by those researchers who 'drink' their research from the same sources, or a community could be formed by those researchers who are cited by the same colleagues. Obviously, in this situation, two or more nodes may belong to the same community/cluster even if they are not directly connected by edges.
- The last one is related with the idea of flow. In this case, a group is as good as much information can be moved within it. Although this concept of community is clearly related with the idea of density (since the more relations between the members of a group there are, probably the higher the flow capacity will be), it is important to note that they differ in many respects. In flow problems, the structure and location of the edges can be decisive when we

have to distribute the information (by flow). Obviously, this is not reflected by the density of a group that only counts how many edges there are over the totals without specifying the way in which they are arranged.

In this work, we focus on this last idea of group, trying to identify groups in which the information moved by the flow is important [4]. It is easy to find many examples of this type of graphs in the field of social networks. For example, Twitter is a directed network in which each arrow may indicate the number of messages of i retweeted by j. Scientific reference pages such as Scopus, WOS or Google Academic, are also directed networks in which each arrow may indicate the number of times that author i has been cited by author j. In both cases, the flow measures the influence of i over j. A metro or a road network are also two examples of directed networks whose community structure depends on the flow.

The key is to find the way to incorporate this group definition into community detection problems. The use of fuzzy sets in social network analysis problems, and in particular, in community detection problems, is not new [10,11,13,23,26,27]. Due to the way in which imprecision is modelled, fuzzy sets appear in a natural way when modelling real problems. In this sense, this paper proposes the use of fuzzy measures or capacity measures [7,25] to measure the relative strength of a group, according to the ability of their members to communicate among them. The more flow the members can send, the more cohesive the group will be.

Once the graph and the fuzzy measure are modelled, in this paper we provide a very efficient algorithm that combines the two class of information (the network and the fuzzy measure), allowing us to identify groups in which the idea of flow is considered. The proposed method may also be useful in the size reduction of large scale fuzzy cognitive maps [14], since their structure is a weighted directed digraph.

The rest of the paper is organized as follows. In Sect. 2 we introduce some basic definitions about community detection problems and fuzzy measures background. In Sect. 3 we introduce a new fuzzy measure related to the flow of a directed network. In Sect. 4 we propose an algorithm to deal with community detection problems with fuzzy measures in directed networks. Finally, some conclusions and future research are shown in Sect. 5.

2 Preliminaries

In this Section we introduce several concepts, definitions and algorithms necessary to have a proper understanding of this paper.

2.1 Community Detection Problems in Directed Networks Based on Density

Definition 1. Directed Network [18]. *A directed network is a set of individuals connected together, in which all the edges are directed from one individual to*

another. A directed network is usually represented by a graph $G = (V, E)$*, where* V *is the set of individuals, called nodes or vertices, and* $E = \{(i,j)|\ i,j \in V\}$ *is the set of ordered pairs of* $V \times V$*, which are directed edges connecting pairs of nodes* (i,j)*. Another way to represent directed graphs or networks is by means of its adjacency matrix,* A*, defined as follows:*

$$A_{ij} = \begin{cases} 1 & if \quad (i,j) \in E, \quad \forall i,j \in 1,\ldots,|V| \\ 0 & otherwise \end{cases}$$

where 1 represents the directed edge which connects i *with* j*.*

Then, let us recall the definition of community detection problems. Given a graph, this type of problem consists in finding a 'good' partition for the input set of individuals. The notion of 'good' may be different depending on the interests of each problem. Many measures have been proposed in the literature to quantify the goodness of a partition [16]. One of the most popular is the modularity, introduced by Girvan and Newman [22] for non-directed networks. This measure has been adapted to directed networks.

Definition 2. Directed Modularity $\mathbf{Q_d}$ [1]. *The modularity is a quality function to measure the goodness of a partition. Let* G *be a directed graph and* P *a partition of the nodes. The directed modularity is defined as:*

$$Q_d(G,P) = \frac{1}{m}\sum_{i,j}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right]\delta(c_i,c_j) \tag{1}$$

where $\delta(c_i,c_j)$ *is* 1 *if* i *belongs to the same group than* j*, and 0 otherwise,* m *is the amount of edges, and* k_i^{in} *and* k_i^{out} *are the in/out edges of node* i*.*

There are many methods to deal with community detection problems [3,8,21]. Particularly, we focus on one of the most popular methods: Louvain Algorithm [2]. Because of its effectiveness and speed, it is one of the most used algorithms. It is based on modularity optimization, and works very well in large networks.

2.2 Fuzzy Measures, Directed Fuzzy Graphs, Extended Fuzzy Directed Graphs

Definition 3. Fuzzy Measure [25]. *Given a finite set* V*, a fuzzy measure is a function* $\mu : 2^V \longrightarrow [0,1]$ *that is monotonous (*$\forall A, B \subseteq V$ *such that* $A \subseteq B$*) and normalized (*$\mu(V) = 1$*), and that satisfies the boundary condition (*$\mu(\emptyset) = 0$*).*

A characteristic of fuzzy measures is their k-additivity [12]. Particularly in this paper, we work with 2-additive fuzzy measures, so let us characterize them.

Definition 4. 2 – additive fuzzy measure [12]. *The fuzzy measure* $\mu : 2^V \longrightarrow [0,1]$ *is said to be 2-additive if and only if,* $\forall S \subseteq V$*, it can be written as a linear combination* $\mu(S) = \sum_{i=1}^{n} a_i x_i + \sum_{\{i,j\}\subset A} a_{ij}x_i x_j$*, where* $x_i = 1$ if $i \in S$ and $x_i = 0$ otherwise.

Then, let us recall the notion of extended fuzzy graph, firstly introduced for non-directed networks in [13].

Definition 5. Extended Fuzzy Graph [13]. *Let $G = (V, E)$ be a graph, and let $\mu : 2^V \longrightarrow [0, 1]$ be a fuzzy measure defined over the set of nodes. The triplet $\widetilde{G} = (V, E, \mu)$ obtained from considering together the graph with the fuzzy measure, is called extended fuzzy graph.*

Note that this structure is much more complex than a fuzzy graph [23]. Fuzzy graphs could be somehow seen as weighted graphs, as the only available information is provided by their edges and their membership degree.

3 Fuzzy Measures from a Directed Networks: The Flow Capacity Measure

Classical community detection problems just consider the topological information provided by the adjacency matrix of networks. Other evidences, such as that given by the flow of the graph, have not been previously considered when dealing with this type of problems. Then, in this Section we propose a way to use the flow in community detection problems. To deal with it, we propose the use of a fuzzy measure which models the relative flow, by means of the weight of the edges. This weight represents the different degree of communication ability of each link, something obvious in real-life problems, in which different relations may have different importance. Then, we give a group idea related to the flow.

Definition 6. *Let $G = (V, E)$ be a directed graph, let (i, j) be an edge, and let f_{ij} be the flow between nodes i and j in G [9]. Then, $\forall S \subseteq V$, we define the function: $\mu^F(S) = \frac{\sum_{i,j \in S} f_{ij}}{\sum_{i,j \in V} f_{ij}}$.*

As a capacity measure, μ^F represents the communication capacity within a set of nodes of a directed network.

Proposition 1. *The function μ^F introduced in Definition 6 is a fuzzy measure.*

Proof. We will verify that μ^F meets the points mentioned in Definition 3.

1. $\mu^F(\emptyset) = 0$ Trivial.
2. $\mu^F(V) = 1$ due to normalization.
3. Let $A \subseteq B \subseteq V$. Then,
   $\mu^F(B) = \frac{\sum_{i,j \in B} f_{ij}}{\sum_{i,j \in V} f_{ij}} = \frac{\sum_{i,j \in A} f_{ij} + \sum_{i,j \in B \setminus A} f_{ij}}{\sum_{i,j \in V} f_{ij}} = \frac{\sum_{i,j \in A} f_{ij}}{\sum_{i,j \in V} f_{ij}} + \frac{\sum_{i,j \in B \setminus A} f_{ij}}{\sum_{i,j \in V} f_{ij}} =$
   $\mu^F(A) + \mu^F(B \setminus A) \geq \mu^F(A)$, since for all $i, j \in V$, $f_{ij} \geq 0$.

Proposition 2. *The fuzzy measure μ^F introduced in Definition 6 is a 2-additive fuzzy measure [12].*

Proof. We will verify that $\mu^F(S)$ can be defined as a linear combination:

$$\mu^F(S) = \sum_{i=1}^{n} a_i x_i + \sum_{\{i,j\} \in V} a_{ij} x_i x_j$$

where $x_i = 1$ if i ∈ S and $x_i = 0$ otherwise.

Let us define: $a_i = \frac{f_{ii}}{\sum_{l,m \subseteq V} f_{lm}}$, and $\mathrm{a_{ij}} = \frac{\mathrm{f_{ij}} + \mathrm{f_{ji}}}{\sum_{\mathrm{l,m} \subseteq \mathrm{V}} \mathrm{f_{lm}}}$. Then, we can write:

$$\mu^F(S) = \sum_{i=1}^{n} \frac{f_{ii}}{\sum_{l,m \subseteq V} f_{lm}} x_i + \sum_{\{i,j\} \in V} \frac{f_{ij} + f_{ji}}{\sum_{l,m \subseteq V} f_{lm}} x_i x_j$$

Once we have the fuzzy measure μ^F which models the ability of the flow in a directed network, here we propose to build the graph associated with it, G_{μ^F}. To carry on with it, we work with the interaction index proposed by Grabisch [12]. Let us denote $\mu_S := \mu(S)$.

Definition 7. Interaction Index [12]*: Let V be a finite set, and let μ be a fuzzy measure defined over it. Let $\{i, j\} \in V$. The interaction index introduced by Grabisch, I_{ij} is defined as:*

$$I_{ij} = \sum_{k=0}^{n-2} \zeta_k \sum_{\substack{K \subset V \setminus \{i,j\} \\ |K| = k}} (\mu_{ijK} - \mu_{iK} - \mu_{jK} + \mu_K) \tag{2}$$

where $\zeta_k = \frac{(n-k-2)!k!}{(n-1)!} = \frac{1}{\binom{n-2}{k}(n-1)}$

Given two items i, j, the interaction index related to a fuzzy measure, represents a class of dependency/association in the global capacity. In this way, it is possible to construct a valued graph from a fuzzy measure which defines these dependencies. In [13], this was the way in which the fuzzy measure was taken into account for the community detection problem. We would like to emphasize that the capacity measure will be taken into account in the clustering problem thanks to the interaction index that will force some nodes to be in the same group while others separated.

Proposition 3. *Let $G = (V, E)$ be a directed graph whose related flow function is f, and let μ^F be the fuzzy measure introduced in Definition 6. Then:*

$$I_{ij} = \frac{f_{ij} + f_{ji}}{\sum_{l,m \in V} f_{lm}} \tag{3}$$

Proof. From equation (2), we can rewrite the components of μ^F as:

$$\mu^F_{ijK} = \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{li}+f_{il}}{\sum_{r,m\subseteq V} f_{rm}} + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{lj}+f_{jl}}{\sum_{r,m\subseteq V} f_{rm}} + \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}}$$

$$\mu^F_{iK} = \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{li}+f_{il}}{\sum_{r,m\subseteq V} f_{rm}}$$

$$\mu^F_{jK} = \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{lj}+f_{jl}}{\sum_{r,m\subseteq V} f_{rm}}$$

$$\mu^F_{K} = \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}}$$

Hence, transcribing and reducing:

$$\begin{aligned} I_{ij} = {} & \sum_{k=0}^{n-2} \zeta_k \sum_{\substack{K\subset V\setminus\{i,j\}\\ |K|=k}} \Big[\sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{li}+f_{il}}{\sum_{r,m\subseteq V} f_{rm}} \\ & + \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{lj}+f_{jl}}{\sum_{r,m\subseteq V} f_{rm}} + \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}} - \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} \\ & - \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{li}+f_{il}}{\sum_{r,m\subseteq V} f_{rm}} - \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} - \sum_{\substack{l\in K\\ l\neq i,j}} \frac{f_{lj}+f_{jl}}{\sum_{r,m\subseteq V} f_{rm}} \\ & + \sum_{\substack{l\in K\\ l\neq i,j}} \sum_{\substack{s\in K\\ s\neq i,j}} \frac{f_{ls}}{\sum_{r,m\subseteq V} f_{rm}} \Big] = \sum_{k=0}^{n-2} \zeta_k \sum_{\substack{K\subset V\setminus\{i,j\}\\ |K|=k}} \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}} \\ = {} & \sum_{k=0}^{n-2} \frac{1}{\binom{n-2}{k}(n-1)} \binom{n-2}{k} \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}} = \sum_{k=0}^{n-2} \frac{1}{n-1} \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}} \\ = {} & \frac{f_{ij}+f_{ji}}{\sum_{r,m\subseteq V} f_{rm}} = I_{ij} \end{aligned}$$

In the classical definition of the interaction index, the order of the elements i and j in the pair $\{i,j\}$ has no significance. Then, we propose an adaptation of it, in order to consider those cases in which this order is important.

Definition 8. Directed Interaction Index. *Let $G = (V,E)$ be a directed graph whose related flow function is f. Let μ be a fuzzy measure defined over the set of nodes, V. Given a pair of ordered nodes (i,j), where $i,j \in V$, we define the directed interaction index I^D_{ij} as:*

$$I^D_{ij} = \frac{f_{ij}}{\sum_{l,m\in V} f_{lm}} \tag{4}$$

From previous definition, we can trivially see that $\forall i,j,\ i\neq j,\ I_{ij} = I^D_{ij} + I^D_{ji}$. Then, the adjacency matrix of the G_{μ^F} is the matrix I^D.

Let us illustrate the calculation of μ^F and I^D with a toy example.

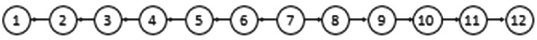

Fig. 1. Directed chain with 12 nodes.

Example 3.1. *We evaluate a simple example of a chain with 12 nodes, as it is drawn in Fig. 1. Let us assume that the weight of all the edges is* 1*. Then, we calculate* μ^F *and* I^D*.*

In this example we calculate the fuzzy measure μ^F *which, as it is shown previously, is 2-additive. Therefore, we only have to calculate it for those subsets of* V *with cardinality one and two. We also calculate the directed interactions,* I^D*. See Fig. 2.*

As it can be seen in matrix I^D*, the node 7 is the only that can communicate with the rest of nodes. At the same time, it is appreciable how this chain is divided by means of its flow values. The nodes 6, 5, 4, 3 and 2 only can communicate with the node with which the related flow reaches the lowest value. In the same way, 8, 9, 10, 11 can connect with the node with which the related flow reaches the highest value. On the other hand, 1 and 12 are isolated. Also, let us observe that in both matrices, the blanks mean 0, in Fig. 2.*

$$\mathrm{A}=\begin{pmatrix}
 & & & & & & & & & & & \\
1 & & & & & & & & & & & \\
 & 1 & & & & & & & & & & \\
 & & 1 & & & & & & & & & \\
 & & & 1 & & & & & & & & \\
 & & & & 1 & & & & & & & \\
 & & & & & 1 & 1 & & & & & \\
 & & & & & & & 1 & & & & \\
 & & & & & & & & 1 & & & \\
 & & & & & & & & & 1 & & \\
 & & & & & & & & & & 1 & \\
 & & & & & & & & & & & 1
\end{pmatrix}
\quad
\mu^F(\{i,j\})=\tfrac{1}{36}\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & & & & & & \\
 & 1 & 1 & 1 & 1 & 1 & & & & & & \\
 & & 1 & 1 & 1 & 1 & & & & & & \\
 & & & 1 & 1 & 1 & & & & & & \\
 & & & & 1 & 1 & & & & & & \\
 & & & & & 1 & & & & & & \\
 & & & & & & 1 & 1 & 1 & 1 & 1 & \\
 & & & & & & & 1 & 1 & 1 & 1 & \\
 & & & & & & & & 1 & 1 & 1 & \\
 & & & & & & & & & 1 & 1 & \\
 & & & & & & & & & & 1 &
\end{pmatrix}
\quad
I^D_{ij}=\tfrac{1}{36}\begin{pmatrix}
 & & & & & & & & & & & \\
1 & & & & & & & & & & & \\
1 & 1 & & & & & & & & & & \\
1 & 1 & 1 & & & & & & & & & \\
1 & 1 & 1 & 1 & & & & & & & & \\
1 & 1 & 1 & 1 & 1 & & & & & & & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \\
 & & & & & & & 1 & 1 & 1 & 1 & \\
 & & & & & & & & 1 & 1 & 1 & \\
 & & & & & & & & & 1 & 1 & \\
 & & & & & & & & & & 1 &
\end{pmatrix}$$

Fig. 2. Adjacency matrix A, μ^F and interaction matrix I^D of directed chain.

4 Community Detection Problems with Capacity Measures in Directed Networks

In the previous Section, we have defined a fuzzy measure that represents the flow capacity of a group in a directed network. In this Section, we will take it into account to find communities in directed networks. The idea of using fuzzy measures in community detection problems was firstly introduced in [13] for non-directed graphs. There it is shown that the original concept of group/community change when a fuzzy measure is also considered, apart from the connections among nodes.

As we have pointed out in the introduction, our aim is to identify groups in which the idea of flow is considered. It is important to note that the modularity measure introduced in [20] for directed networks does not consider the flow. Then, as a natural consequence, any algorithm based on modularity optimization, will not be suitable for searching communities based on the flow. In order to show this fact with more emphasize, we propose another expression for the modularity formula.

Let $G = (V, E)$ be a directed graph, let $S \subseteq V$, and let $P = \{C_1, \ldots, C_L\}$ be a partition of the set of nodes. Here we introduce some notation:

- $K_S^{in} = \sum_{i \in S} k_i^{in}$, the number of links that goes to any node of S.
- $K_S^{out} = \sum_{i \in S} k_i^{out}$, the number of links that goes out from any node of S.
- $m_S = \sum_{i,j \in S} A_{ij}$, the number of links among the members of S.

Now, we can consider another expression of directed modularity [1] introduced by Newman for a given partition $P = \{C_1, \ldots, C_L\}$.

$$Q_d(G, P) = \frac{1}{m} \sum_{l=1}^{L} \sum_{i,j \in C_l} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] = \frac{1}{m} \sum_{l=1}^{L} \left[m_{C_l} - \frac{K_{C_l}^{in} K_{C_l}^{out}}{m} \right].$$

From previous expression we can see that, fixed the edges in a group, the distribution and localization have not any impact in the modularity measure. The reason is that the important things of the measure are: the number of links inside the group, the number of links that goes from one element of the group to another (of the group or not), and the number of links that influences any member of the group (the origin of each link has no significance). In following example, we show it in detail.

Example 4.1. *Let us consider three directed graphs,* $G_i = (V_i, E_i)$ *for* $i = 1, \ldots, 3$, *where* $|V_1| = |V_2| = |V_3| = 6$ *and* $|E_1| = |E_2| = |E_3| = 5$.

Let us denote $V_1 = \{1, 2, \ldots, 6\}$, $V_2 = \{7, 8, \ldots, 12\}$, *and* $V_3 = \{13, 14, \ldots, 18\}$. *We assume that the graphs* G_1 *and* G_3 *are two directed stars (with hubs* 1 *and* 13*) and let us suppose that* G_2 *is a 6-directed chain.*

Let $E_1 = \{(1,2), (1,3), (1,4), (1,5), (1,6)\}$; $E_2 = \{(7,8), (8,9), (9,10), (10,11), (11,12)\}$ *and* $E_3 = \{(13,14), (15,13), (13,16), (13,17), (13,18)\}$ *be the sets of edges of these graphs, respectively.*

Then, we present the following networks built from the aggregation of two of the previous graphs. $G_{12} = (V_1 \cup V_2, E_1 \cup E_2 \cup \{(6,7)\})$; $G_{32} = (V_3 \cup V_2, E_3 \cup E_2 \cup \{(18,7)\})$; $G_{13} = (V_1 \cup V_3, E_1 \cup E_3 \cup \{(6,13)\})$ *(Fig. 3).*

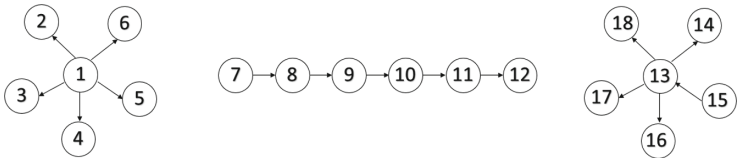

Fig. 3. Two directed stars and a 6-directed chain.

If we break these networks as $P_{12} = \{V_1, V_2\}$; $P_{32} = \{V_3, V_2\}$, $P_{13} = \{V_1, V_3\}$ *the modularity of each graph is:*

$$\begin{aligned} Q_d(G_{12}, P_{12}) &= \frac{1}{m}\sum\nolimits_{i,j\in V_1}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] + \frac{1}{m}\sum\nolimits_{i,j\in V_2}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] \\ &= \frac{1}{11}\left[5 - \frac{30}{11}\right] + \frac{1}{11}\left[5 - \frac{30}{11}\right] = \frac{25}{121} + \frac{25}{121} \\ Q_d(G_{32}, P_{32}) &= \frac{1}{m}\sum\nolimits_{i,j\in V_3}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] + \frac{1}{m}\sum\nolimits_{i,j\in V_2}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] \\ &= \frac{1}{11}\left[5 - \frac{30}{11}\right] + \frac{1}{11}\left[5 - \frac{30}{11}\right] = \frac{25}{121} + \frac{25}{121} \\ Q_d(G_{13}, P_{13}) &= \frac{1}{m}\sum\nolimits_{i,j\in V_1}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] + \frac{1}{m}\sum\nolimits_{i,j\in V_3}\left[A_{ij} - \frac{k_i^{in}k_j^{out}}{m}\right] \\ &= \frac{1}{11}\left[5 - \frac{30}{11}\right] + \frac{1}{11}\left[5 - \frac{30}{11}\right] = \frac{25}{121} + \frac{25}{121} \end{aligned}$$

Therefore, taking into account the inconveniences we found to get a 'good' partition with the current method, we introduce a capacity measure algorithm based on flow which will find clusters with maximum flow.

We propose a modification of Directed Louvain Algorithm [5,17] to work with extended fuzzy directed graphs denoted as Flow Capacity Louvain.

Let us define some concepts related to the Algorithm 1:

- $\Delta Q_i^d(j)$ is the increase in modularity when node j is incorporated into the community of i.
- A is the adjacency matrix, which has to guarantee the connections among the nodes.
- I^D is the directed interaction matrix.
- $\alpha \in [0, 1]$ parameter of importance [13] which assigns a weight to each part of the extended fuzzy directed graph.
- $M = \alpha A + (1 - \alpha)\, I^D$ is the matrix in which we search the partition, by maximizing its modularity.

Let us illustrate the performance of Flow Capacity Louvain Algorithm.

Algorithm 1. *Flow Capacity Louvain* **input**=A **output**=P

Phase 1.
$o = permutation(V)$
Let each node of the graph be an isolated community
while There is some change in modularity optimization **do**
  According to the order given by o, let i be the corresponding element.
  Then, find all out-edges j and in-edges j of i in A
  Calculate $\Delta Q_i^d(j)$ in matrix $M = \alpha A + (1-\alpha)I^D$
  Let j^* be the node for which ΔQ_i^d is maximum
  if $\Delta Q_i^d(j^*) > 0$ **then**
    Move node i to the community to which j^* belongs
  else
    i remains in its community
  end if
end while
Phase 1 Ends
Phase 2.
A^* is the aggregated matrix obtained from A, whose nodes are the communities found in Phase 1
M^* is the aggregated matrix obtained from M, whose nodes are the communities found in Phase 1
While there is some change, apply Flow Capacity Louvain Algorithm, considering matrix A^* to find nodes and M^* to modularity optimization
Phase 2 Ends

Example 4.2. *Let us recall the graph introduced in Example 3.1. The partition* P_1 *obtained with the directed Louvain's algorithm divides this chain in three parts. The central cut* $\{5,6,7,8\}$ *has a bad behavior on flow (the modularity of* I^D *is not good). However, the partition* P_2 *obtained with Flow Capacity Louvain Algorithm (with* $\alpha = 0.5$*), defines 2 slices, both with good behavior on flow. All results can be seen in the Table 1.*

Table 1. Modularity of several partitions of the chain.

	Clustering classification according to chosen cut		
Directed Louvain	$P_1 = \{\{1,2,3,4\};\{5,6,7,8\};$ $\{9,10,11,12\}\}$	$Q_d(A,P_1) =$ ***0.496***	$Q_d(I^D,P_1) =$ ***0.222***
Flow capacity	$P_2 = \{\{1,2,3,4,5,6\};$ $\{7,8,9,10,11,12\}\}$	$Q_d(A,P_2) =$ ***0.413***	$Q_d(I^D,P_2) =$ ***0.347***

Example 4.3. *Let us consider three directed circles* $(1,\ldots,6)$, $(1',\ldots,6')$ *and* $(1'',\ldots,6'')$. *The graph of Fig. 4 is obtained by connecting each vertex of the first circle with its corresponding node of the second circle, and each vertex of the third circle with its corresponding node of the second circle. The interaction matrix* I^D *and the adjacency matrix* A *are represented in the Fig. 5. Let us consider that the weight of all the edges is* 1. *This structure can approach a wheel.*

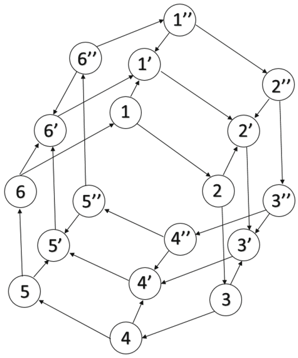

Fig. 4. Directed wheel with 18 nodes.

$$
A=\begin{pmatrix}
 & 1 & & & & & 1 & & & & & & & & & & & \\
 & & 1 & & & & & 1 & & & & & & & & & & \\
 & & & 1 & & & & & 1 & & & & & & & & & \\
 & & & & 1 & & & & & 1 & & & & & & & & \\
 & & & & & 1 & & & & & 1 & & & & & & & \\
1 & & & & & & & & & & & 1 & & & & & & \\
 & & & & & & & 1 & & & & & & & & & & \\
 & & & & & & & & 1 & & & & & & & & & \\
 & & & & & & & & & 1 & & & & & & & & \\
 & & & & & & & & & & 1 & & & & & & & \\
 & & & & & & & & & & & 1 & & & & & & \\
 & & & & & & 1 & & & & & & & & & & & \\
 & & & & & & 1 & & & & & & & 1 & & & & \\
 & & & & & & & 1 & & & & & & & 1 & & & \\
 & & & & & & & & 1 & & & & & & & 1 & & \\
 & & & & & & & & & 1 & & & & & & & 1 & \\
 & & & & & & & & & & 1 & & & & & & & 1 \\
 & & & & & & & & & & & 1 & 1 & & & & &
\end{pmatrix}
\qquad
I^D=\frac{1}{252}\begin{pmatrix}
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
1&1&1&1&1&1&2&2&2&2&2&2& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &1&1&1&1&1&1& & & & & & \\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1\\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1\\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1\\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1\\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1\\
 & & & & & &2&2&2&2&2&2&1&1&1&1&1&1
\end{pmatrix}
$$

Fig. 5. Adjacency matrix A and Directed Interaction matrix I^D of directed wheel.

On this occasion, considering Louvain Algorithm [2], there are several partitions which maximize the modularity, cutting the wheel into three triangular zones. One of these partitions with maximum modularity in A is the cut P_1. In contrast, the modularity related to the flow (I^D) which is associated to P_1, is 0. Nevertheless, considering Flow Capacity Louvain Algorithm (with $\alpha = 0.5$), the obtained partition P_2 guarantees a high but not maximum modularity in A, and a maximum modularity in I^D. All the results can be seen in Table 2.

In this example, it is clearly seen how maximizing directed modularity does not achieve groups that maximize flow.

Let us remark that, in all these examples, we have assumed that the weight of all the edges is 1. Nevertheless, if we incorporate different weights, our algorithm will work correctly since Louvain Algorithm does.

Table 2. Modularity of several partitions of the wheel.

	Clustering classification according to chosen cut		
Directed Louvain	$P_1 = \{\{1, 1', 1'', 2, 2, 2''\}; \{3, 3', 3'', 4, 4', 4''\} \{5, 5', 5'', 6, 6', 6''\}\}$	$Q_d(A, P_1) = 0.367$	$Q_d(I^D, P_1) = 0$
Flow Capacity	$P_2 = \{\{1, 2, 3, 4, 5, 6\}; \{1', 2', 3', 4', 5', 6'\} \{1'', 2'', 3'', 4'', 5'', 6''\}\}$	$Q_d(A, P_2) = 0.320$	$Q_d(I^D, P_2) = 0.204$

Moreover, we would like to mention that Flow Capacity Louvain Algorithm complexity will be the highest between maximum flow among all pairs of nodes and Louvain Algorithm complexity [2,6].

5 Conclusions

Contrary to the non-directed networks case, the idea of a group/community in directed networks allows different interpretations depending on the objective of the clustering problem. In general, most community detection algorithms for directed networks focus on the idea of *group by density*, or in the idea of *random walker group*. Few works have been developed to identify groups in which the idea of the flow group is considered. In this paper, we deal with the flow community detection problem that tries to find communities in which not only it is important to potentiate communities with many connections among its members, but also it is important to maintain together these connections that allow the flow of information among its members. It is important to note that the measure of modularity proposed by Newman for directed networks finds communities with high density but not necessary well connected, regarding the flow. As we show with some examples in this work, modularity (and, as a consequence any optimization algorithm based on it), does not distinguish between different situations in which it is necessary to add some information to identify communities.

In order to take into account the flow capacity of a group, we incorporate to the community detection problem a 2-additive fuzzy measure that represents the relative flow capacity of each set of nodes. Then, following a similar methodology as that introduced in [13], we propose a modification of the Directed Louvain Algorithm [5,17] in order to incorporate the information provided by a fuzzy measure to the community detection algorithm in directed networks. Our proposal, Flow Capacity Louvain Algorithm, can consider, analyze and apply the information defined by a fuzzy measure when finding a partition in a directed network. Particularly, we propose to consider the fuzzy measure μ^F introduced in Sect. 3. Under the assumption of the new group definition based on the flow, we show that this algorithm provides very good results. As further work, we will develop an experimental study to test the efficiency of the algorithm here proposed, as well as an analysis of the processing time and memory usage of it. We will also work in some computational results, considering several benchmark models apart from the examples that we have included in this paper.

References

1. Arenas, A., Duch, J., Fernández, A., Gómez, S.: Size reduction of complex networks preserving modularity. New J. Phys. **9**(6), 176 (2007)
2. Blondel, V., Guillaume, J., Lambiotte, R., Lefevre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.-Theory Exp. (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
3. Bollobás, B.: Modern Graph Theory, pp. 215–252. Springer, New York (1998). https://doi.org/10.1007/978-1-4612-0619-4
4. Borgatti, S.P.: Centrality and network flow. Soc. Netw. **27**(1), 55–71 (2005)
5. Dugué, N., Perez, A.: Directed Louvain: maximizing modularity in directed networks (2015)
6. Edmonds, J., Karp, R.M.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM (JACM) **19**(2), 248–264 (1972)
7. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
8. Girvan, M., Newman, M.: Community structure in social and biological networks. Proc. Nat. Acad. Sci. **99**(12), 7821–7826 (2002)
9. Goldberg, A.V., Tarjan, R.E.: A new approach to the maximum-flow problem. J. ACM (JACM) **35**(4), 921–940 (1988)
10. Gómez, D., Rodríguez, J., Yáñez, J., Montero, J.: A new modularity measure for Fuzzy Community detection problems based on overlap and grouping functions. Int. J. Approx. Reason. **74**, 88–107 (2016)
11. Gómez, D., Zarrazola, E., Yáñez, J., Montero, J.: A divide-and-link algorithm for hierarchical clustering in networks. Inf. Sci. **316**, 308–328 (1997)
12. Grabisch, M.: K-order additive discrete fuzzy measures and their representation. Fuzzy Sets Syst. **92**(2), 167–189 (1997)
13. Gutiérrez, I., Gómez, D., Castro, J., Espínola, R.: A new community detection algorithm based on fuzzy measures. In: Kahraman, C., Cebi, S., Cevik Onar, S., Oztaysi, B., Tolga, A.C., Sari, I.U. (eds.) INFUS 2019. AISC, vol. 1029, pp. 133–140. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-23756-1_18
14. Kosko, B., et al.: Fuzzy cognitive maps. Int. J. Man Mach. Stud. **24**(1), 65–75 (1986)
15. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys. Rev. E **80**(1), 016118 (2009)
16. Li, H., Xiang, J.: Explore of the fuzzy community structure integrating the directed line graph and likelihood optimization. J. Intell. Fuzzy Syst. **32**(6), 4503–4511 (2017)
17. Li, L., He, X., Yan, G.: Improved Louvain method for directed networks. In: Shi, Z., Mercier-Laurent, E., Li, J. (eds.) IIP 2018. IAICT, vol. 538, pp. 192–203. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00828-4_20
18. Malliaros, F., Vazirgiannis, M.: Clustering and community detection in directed networks: a survey. Phys. Rep. **533**(4), 95–142 (2013)
19. Masuda, N., Porter, M., Lambiotte, R.: Random walks and diffusion on networks (vol 716, pg 1, 2017). Phys. Rp-Rw Sect. Phys. Lett. **745**, 96 (2018)
20. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(3), 036104 (2006)
21. Newman, M.: Community detection and graph partitioning. EPL (Europhy. Lett.) **103**(2), 28003 (2013)

22. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. **69**, 026113 (2004)
23. Rosenfeld, A.: Fuzzy graphs. Fuzzy Sets Appl. 77–95 (1975)
24. Rosvall, M., Bergstrom, C.: Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS One **6**(4), e18209 (2011)
25. Sugeno, M.: Fuzzy measures and fuzzy integrals–a survey. In: Readings in Fuzzy Sets for Intelligent Systems, pp. 251–257. Elsevier (1993)
26. Wu, T., Liu, X., Liu, F.: An interval type-2 fuzzy TOPSIS model for large scale group decision making problems with social network information. Inf. Sci. **432**, 392–410 (2018)
27. Zhang, D., Xie, F., Zhang, Y., Dong, F., Hirota, K.: Fuzzy analysis of community detection in complex networks. Phys. A **389**(22), 5319–5327 (2010)