

---

# DYNCOMM R PACKAGE - DYNAMIC COMMUNITY DETECTION FOR EVOLVING NETWORKS

---

**Rui Portocarrero Sarmento**  
PRODEI - Faculty of Engineering, UP  
Rua Dr. Roberto Frias, s/n  
4200-465 Porto, Portugal  
mail@ruisarmento.com

**Luís Lemos**  
FEUP - Faculty of Engineering, UP  
Rua Dr. Roberto Frias, s/n  
4200-465 Porto, Portugal

**Mário Cordeiro**  
PRODEI - Faculty of Engineering, UP  
Rua Dr. Roberto Frias, s/n  
4200-465 Porto, Portugal

**Giulio Rossetti**  
ISTI-CNR  
Via Giuseppe Moruzzi, 1,  
56127 Pisa PI, Italy  
giulio.rossetti@isti.cnr.it

**Douglas Cardoso**  
CEFET-RJ  
Petrópolis, Brazil  
douglas.cardoso@cefet-rj.br

May 7, 2019

## ABSTRACT

Nowadays, the analysis of dynamics in networks represents a great deal in the Social Network Analysis research area. To support students, teachers, developers, and researchers in this work we introduce a novel R package, namely DynComm. It is designed to be a multi-language package, that can be used for community detection and analysis on dynamic networks. The package introduces interfaces to facilitate further developments and the addition of new and future developed algorithms to deal with community detection in evolving networks. This new package has the goal of abstracting the programmatic interface of the algorithms, whether they are written in R or other languages, and expose them as functions in R.

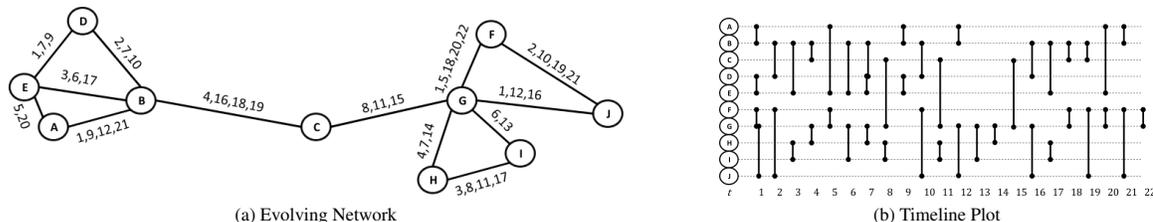
**Keywords** DynComm R Package · R language · Community Detection · Dynamic/Evolving Networks · R Packages

## 1 Introduction

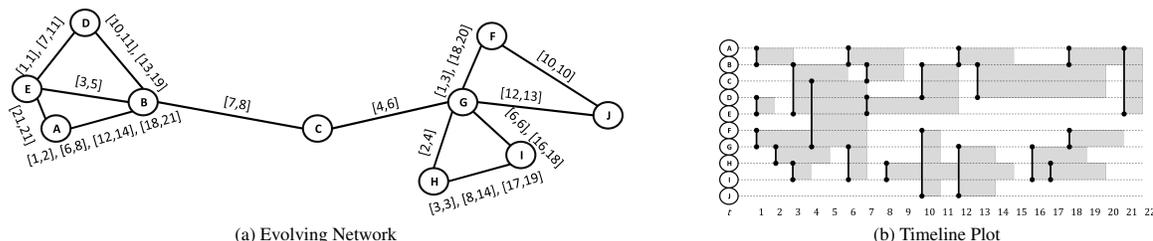
Community Detection in Social Network Analysis (SNA) is a critical research area in an enormous amount of unrelated areas. From Psychology to Physics, community detection in SNA is used to find an agglomeration of objects of study in a graph. We have witnessed the development of an abundance of algorithms specifically designed to identify communities in graphs of every size, from small graphs with some dozens of vertices and edges, to large or very large graphs with millions of vertices and billions of edges.

Very recently, with the growing popularity of SNA, researchers have been migrating concepts and some algorithms to stream approaches. This is even more important with the appearance of sources of data that are streamable, such as social networks like Facebook and Twitter, where information arrives as a flow of discrete events that, usually, have a limited existence through time.

Available since some years ago, a variety of packages in languages as Python or R have been developed to cope with the need for analysis of communities and social networks. Nonetheless, few or no packages that deal with community detection in evolving networks are available right now, in R-CRAN. Thus, this is the right time to develop a way to provide researchers with a package that fulfills the need to explore network streams, particularly community detection of evolving networks. This poses a challenge since the adaptation of algorithms is not an easy task, and sometimes even impossible due to restrictions in the architecture of the static algorithm.



**Figure 1.** Example of contact Evolving Network: Figure 1a shows a labelled aggregate network where the labels denote the times of contact, and Figure 1b shows a time-line plot, where each of the lines corresponds to one vertex and time runs from left to right



**Figure 2.** Example of interval Evolving Network: Figure 2a shows the labelled aggregate network where the labels denote the time interval of the relation, and Figure 2b shows a time-line plot, where each of the lines corresponds to one vertex and grey zones the time duration between two edges

We tried to develop a framework that we believe will help future developments in this area to be included in the package, with the least amount of effort for the new algorithms' authors.

The paper is organized as follows. In Section 2, we introduce some concepts related to evolving networks. Then, in section 3, we explore the developed package and list its features, in the present version of development. In section 3.1.1, we explain the developed R interface, in more detail. Then, in section 3.2.1, we explain the steps needed to add a new algorithm developed in other languages. Finally, in section 4, we conclude this document by suggesting further developments of the DynComm R package <sup>1</sup>.

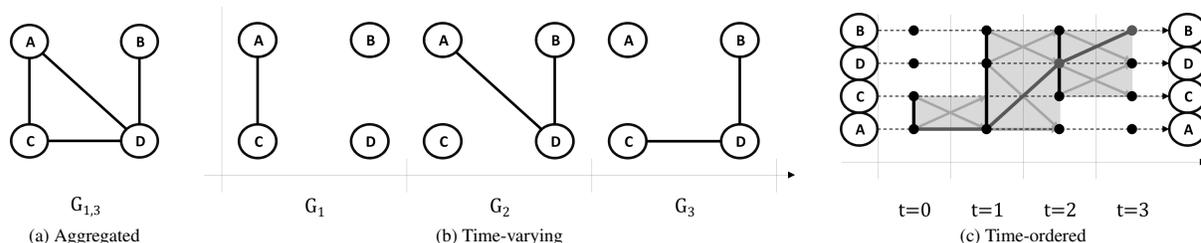
## 2 From Static to Evolving Networks

From (Cordeiro et al., 2018), Figure 1 shows the example of a contact Evolving Network with instantaneous interactions between vertices. When the interaction between network peers has a time duration we are in the presence of interval Evolving Networks as shown in Figure 2b. Assuming that the time  $T$  during which a network is observed is finite we can consider the start point  $t_{start} = 0$  and the end time as  $t_{end} = T$ . A dynamic network graph  $G_{0,T}^D(V, E_{0,T})$  on a time interval  $[0, T[$  consists of a set of vertices  $V$  and a set of temporal edges  $E_{0,T}$ . The evolving network is a set of graphs across the time axis within discrete time points  $t_1, t_2, \dots, t_{n-1}, t_n$ . At time point  $t_n$  is observed a graph instance  $G(V_n, E_n)$  also denoted as  $G_n$  where  $E_n$  is the set of temporal edges  $(u, v)_{t_n} \in E_{0,T}$  at time point  $t_n$  with edges between vertices  $u$  and  $v$  on time interval  $t_n = [t_{n_{begin}}, t_{n_{end}}]$  such that  $t_{n_{begin}} \leq T$  and  $t_{n_{end}} \geq t_{n_{begin}} \geq 0$ . Examples of network changes that may occur between two time points  $t_{n-1}$  and  $t_n$  are the addition of new edges, i.e.:  $E_n \supset E_{n-1}$ , and the appearance of additional vertices, i.e.:  $V_n \supset V_{n-1}$ .

### 2.1 Models of Temporal Representation

Figure 3 presents the concept of a time-ordered graph for an example network for the time interval  $[0, 3]$ . Figure 3a shows all the time intervals aggregated into a single graph  $G_{1,3}$ . The discretization of the network by converting the temporal information into a sequence of  $n$  snapshots is presented in Figure 3b. In this example the evolving network is represented as a series of static networks  $G_1, G_2, \dots, G_n$ . The time-ordered graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  of Figure 3c assumes that at each time step, a message can be delivered along a single edge. In the example of Figure 3c we show the

<sup>1</sup>Available Code at <https://github.com/softskillsgroup/DynComm-R-package>.

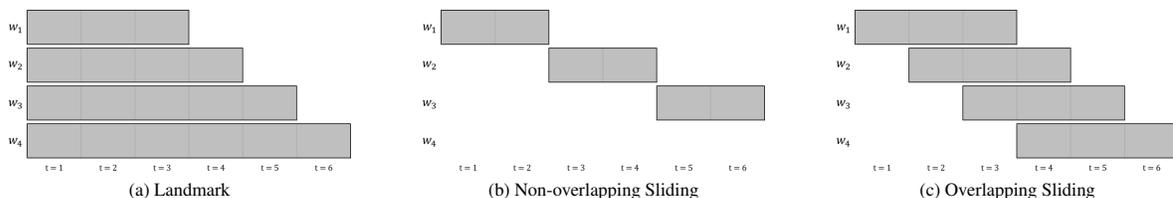


**Figure 3.** Comparison of aggregated representation (Figure 3a) and time series representation (Figure 3b). The corresponding time ordered graph  $\mathcal{G}$  is presented for the interval  $[0, 3]$  (Figure 3c).

temporal shortest path from vertex  $u = A$  to vertex  $v = B$ . The temporal shortest path from  $A$  to  $B$  in the interval  $[0, 3]$  is  $A_0 \rightarrow A_1 \rightarrow D_2 \rightarrow B_3$ . The time-ordered graph of (Kim and Anderson, 2012) is the model used in the rest of this document and package.

## 2.2 Landmark vs Sliding Windows

When the temporal dimension is added to the analysis of networks, methodologies relating to the strategy to deal with knowledge that is being analyzed vary. Figure 4 shows 3 kinds of graph knowledge windowing ways. **Landmark windows** (Gehrke et al., 2001) comprehend all the info from a particular purpose in time, up to the present moment. within the Landmark window, the model is initialized in an determined point in time, i.e., the landmark that marks the start of the window. In ordered snapshots, the info window grows to think about all the data seen up to now, since the landmark start. **Sliding windows**, from other point of view, are appropriate when we are not inquisitive about computing statistics over all the past, solely over the recent past (Gama, 2010). (Datar et al., 2002) incorporates a forgetting mechanism, does not take into account all the data falling outside the window, by keeping solely the newest data within the window. These windows can be defined regarding length in two distinct ways, the time-based length and the sequence-based (Babcock et al., 2002b,a). Sequence-based models, wherever the dimensions of the window are, is set relating to the amount of observations. In Timestamp-based models, the other type of window generation, the dimensions of the window is outlined regarding time sample length. A timestamp window of size  $t$  consists of all event elements whose timestamp is within a time interval  $t$  since the beginning of the data processing, or since the beginning of the current period of processed data.

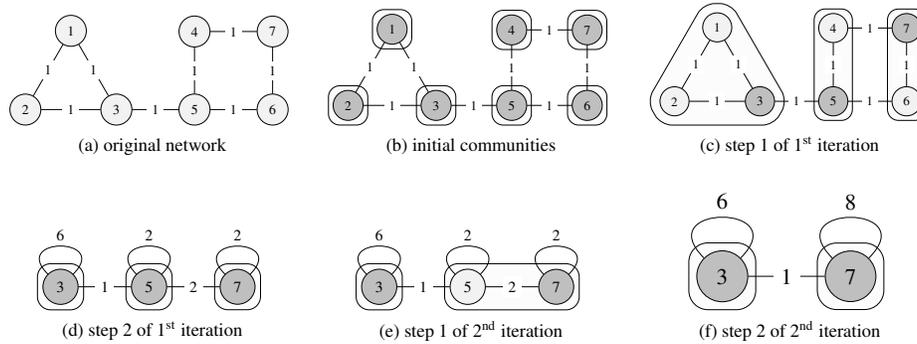


**Figure 4.** Types of Data Windows: Landmark Window (Figure 4a) Non-overlapping Sliding Window (Figure 4b) and Overlapping Sliding Window (Figure 4c)

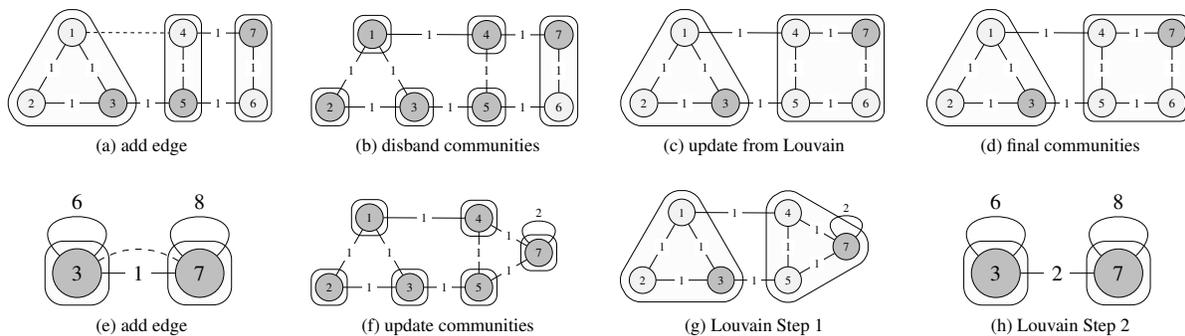
The implementation currently available in our package is explained by 4b.

## 2.3 Dynamic Community Detection

As a consequence of both global and local heterogeneity of edge distribution in a graph, specific regions of a graph evidence high concentration of edges within particular regions, called *communities*, whereas inter regions have low concentrations of edges. In the context of networks, these occurrences of groups of vertices in a network that are more densely connected internally than with the rest of the network is called *community structure*. Also known as *modules* or *clusters*, communities can, therefore, be straightforwardly defined as groups of similar vertices. A complete definition using the concept of density can be the following: communities can be understood as densely connected groups of vertices in the network, with sparser connections between them.



**Figure 5.** Example of an agglomerative community detection algorithm. In this case the original Louvain (Blondel et al., 2008) with all algorithm steps.



**Figure 6.** Example of (Cordeiro et al., 2016) Dynamic Louvain for the addition of a Cross Community Edge (1 — 4). Top figures show the lower-level network. At the bottom, are shown, the corresponding upper-level network with aggregated communities.

### 2.3.1 Finding Communities in Static Networks

A greedy algorithm based on modularity optimization has been introduced by (Blondel et al., 2008) where initially all vertices of the graph are put in different communities (Figure 5b). The first step consists of a sequential sweep over all vertices, for each of the neighbors picks the community that yields the largest increase of modularity (Figure 5c). At the end of the sweep, one obtains the first level partition. In the second step, communities are replaced by super vertices, and weight of the edge between the super vertices is the sum of the weights of the edges between the represented communities at the lower level (Figure 5d). The two steps of the algorithm are then repeated, yielding new hierarchical levels and supergraphs (Figure 5f).

### 2.3.2 Finding Communities in Dynamic Networks

When discussing methods for finding communities in dynamic networks the division of methods for slowly evolving networks and streaming networks is consensual (Aggarwal and Subbian, 2014). In the following section, algorithms for both scenarios will be presented and analyzed

**Slowly Evolving Networks** (Cordeiro et al., 2016) presented a modularity-based dynamic community detection algorithm. It is a modification of the original Louvain method where dynamically added and removed vertices and edges only affect their related communities. In each iteration, all the communities that were not affected by modifications to the network maintain unchanged. By reusing community structure obtained by previous iterations, the local modularity optimization step operates in smaller networks. Thus, only affected communities are disbanded to their origin. Compared with the original algorithm (Figure 6), the stability of communities is also an improvement. When compared with the original algorithm, with that algorithm run several times, the results in changes on communities or vertex drift from one community to another are easier to follow, when presented by the dynamic and incremental algorithm. This is due to the fact that only parts of the network change during iterations, the non-determinism of

the algorithm will have reduced effect on the community assignment, providing better community stability than its counterparts.

---

**Algorithm 1** Dynamic Community Detection Algorithm
 

---

```

1:  $V \leftarrow \{u_1, u_2, \dots, u_v\}$ ,  $E \leftarrow \{(i_1, j_1), (i_2, j_2), \dots, (i_e, j_e)\}$ 
2:  $A \leftarrow \text{array}\{(i_1, j_1), \dots, (i_m, j_m)\}$ 
3:  $R \leftarrow \text{array}\{(i_1, j_1), \dots, (i_n, j_n)\}$ 
4: procedure MAIN( $G \leftarrow (V, E), A, R$ )
5:    $\mathcal{C}_{ll} \leftarrow \{C_1, C_2, \dots, C_n\}$ ,  $\mathcal{C}_{ul} \leftarrow \{\}$ ,  $\mathcal{C}_{aux} \leftarrow \mathcal{C}_{ll}$ 
6:   INITPARTITION( $\mathcal{C}_{aux}$ )
7:    $mod \leftarrow \text{MODULARITY}(\mathcal{C}_{aux})$ ,  $old\_mod \leftarrow 0$ 
8:    $m \leftarrow 1$ ,  $n \leftarrow 1$ 
9:   while ( $mod \geq old\_mod \vee m \leq |A| \vee v \leq |R|$ ) do
10:     $\mathcal{C}_{aux} \leftarrow \text{ONELEVEL}(\mathcal{C}_{aux})$ 
11:     $\langle n, c \rangle \leftarrow \text{COMMUNITYCHANGEDVERTICES}(\mathcal{C}_{ll}, \mathcal{C}_{aux})$ 
12:     $\mathcal{C}_{ll} \leftarrow \text{UPDATECOMMUNITIES}(\mathcal{C}_{ll}, n, c)$ 
13:     $old\_mod \leftarrow mod$ ,  $mod \leftarrow \text{MODULARITY}(\mathcal{C}_{ll})$ 
14:     $\mathcal{C}_{ul} \leftarrow \text{PARTITIONTOGRAPH}(\mathcal{C}_{ll})$ 
15:    if  $m \leq |A|$  then
16:       $\langle src, dest \rangle \leftarrow A[m]$ 
17:       $a_{vertices} \leftarrow \text{AFFECTEDBYADDITION}(src, dest, \mathcal{C}_{ll})$ 
18:       $\mathcal{C}_{ll} \leftarrow \text{ADDEDGE}(src, dest, \mathcal{C}_{ll})$ 
19:       $\mathcal{C}_{ll} \leftarrow \text{DISBANDCOMMUNITIES}(\mathcal{C}_{ll}, a_{vertices})$ 
20:       $\mathcal{C}_{ul} \leftarrow \text{SYNCCOMMUNITIES}(\mathcal{C}_{ll}, \mathcal{C}_{ul}, a_{vertices})$ 
21:    end if
22:    if  $n \leq |R|$  then
23:       $\langle src, dest \rangle \leftarrow R[n]$ 
24:       $a_{vertices} \leftarrow \text{AFFECTEDBYREMOVAL}(src, dest, \mathcal{C}_{ll})$ 
25:       $\mathcal{C}_{ll} \leftarrow \text{REMOVEEDGE}(src, dest, \mathcal{C}_{ll})$ 
26:       $\mathcal{C}_{ll} \leftarrow \text{DISBANDCOMMUNITIES}(\mathcal{C}_{ll}, a_{vertices})$ 
27:       $\mathcal{C}_{ul} \leftarrow \text{SYNCCOMMUNITIES}(\mathcal{C}_{ll}, \mathcal{C}_{ul}, a_{vertices})$ 
28:    end if
29:     $\mathcal{C}_{aux} \leftarrow \mathcal{C}_{ul}$ ,  $m \leftarrow m + 1$ ,  $n \leftarrow n + 1$ 
30:  end while
31: end procedure

```

---

**Streaming Networks** Streaming graph algorithms are essential to perform community detection with high frequency data and large or very large networks. In streaming scenarios, the ability to perform deletion of edges in community detection algorithms is important. In short, as discussed in Section 2.2, this will dictate if the method of analysis is to be performed over a sliding window of edges, and therefore edges are deleted from the tail end of the sliding window, or over a landmark window, in case there is no possibility to delete or forget old edges. Several methods were proposed for dynamic community discovery in graph streams. (Wang et al., 2013) motivated by the variability of the underlying social behavior of individuals over different graph regions modeled the problem according to the so-called *local heterogeneity*, where a Local Weighted-Edge-based Pattern (LWEP) summary is efficiently maintained and used afterward to cluster the graph stream and perform dynamic community detection in weighted graph streams. Taking an almost linear time, (Raghavan et al., 2007) investigated and a simple label propagation algorithm that uses the network structure alone as its guide and requires neither optimization of a predefined objective function nor prior information about the communities. By analyzing the problem of real-time community detection in large networks and having by baseline the algorithm proposed by (Raghavan et al., 2007) with linear time- $O(m)$  on a network with  $m$  edges-label propagation, or “epidemic” community detection, (Leung et al., 2009) proposed a method with near linear time community detection in graphs. They identified the characteristics and drawbacks of the base (Raghavan et al., 2007) algorithm and extended it by incorporating different heuristics to facilitate reliable and multifunctional real-time community detection. (Yun et al., 2014) proposed two efficient streaming memory-limited clustering algorithms for community detection based on spectral methods. (Yun and Proutière, 2014) proposed community detection via random and adaptive sampling. (Sariyüce et al., 2016) proposed SONIC, a find-and-merge type of overlapping community detection algorithm that can efficiently handle streaming updates. Recently, (Hollocou et al., 2017) proposed SCODA, a linear streaming algorithm for community detection in very large networks.

### 2.3.3 Density Optimization

Modularity-based algorithms used for community detection have been increasing in recent years. Modularity and its application have been generating controversy since some authors argue it is not a metric without disadvantages. It has been shown that algorithms that use modularity to detect communities suffer a resolution limit and, therefore, it is unable to identify small communities in some situations. In this function of this package, we try to apply a density optimization of communities found by the available algorithms (Sarmiento, 2019). We introduce a metric we call ADC (Average Density per Community); we use this metric to prove our optimization provides improvements to the community density obtained with benchmark algorithms. The results of the optimization algorithm proved to be interesting.

Several developments were made to test the hypothesis. An algorithm was developed, and a metric is introduced in the following sections.

**Average Density per Community (ADC) measure** Average Density per Community (ADC) is the measure that is used to compare the algorithm results and is given by the following formula:

$$ADC = \frac{1}{n_C} \sum_{C_i=1}^n Density(C_i)$$

where  $n_C$  is the number of communities identified in the graph,  $Density(C_i)$  is the density of each community  $C_i$ .

**Optimization Algorithm** Algorithm 2 provides the sequence of tasks we are doing to test the hypothesis. We start by using the results of a community detection algorithm. Then, we try to discover if the communities can be disbanded in smaller communities. These smaller communities are strongly connected components, i.e., groups of vertices with higher density. Then, if the average community density of the disbanded communities is higher than the original community the disbanding is indeed executed. If not, the community founded by the benchmark algorithm is not disbanded and maintains its original id.

---

#### Algorithm 2 Algorithm Pseudo-Code for Optimization of Community Density

---

**Input:** *Communities\_Data* ▷ Vertex List and their Community  
**Output:** *Community\_Results* ▷ New Community Structure

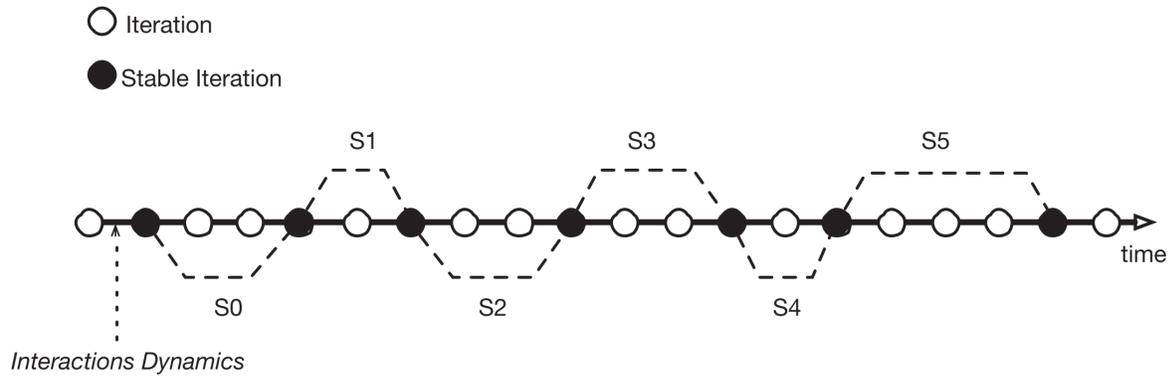
- 1: **while** not at the end of *Original\_Communities* list **do**
- 2:   **if**  $n_{components} > 1$  **then** ▷ If community has more than 1 component
- 3:      $SCC \leftarrow$  STRONG\_CONNECTED\_COMPONENTS\_OF\_COMMUNITY(*Community<sub>i</sub>*)
- 4:      $mdc \leftarrow$  MEAN\_DENSITY\_OF\_COMPONENTS(*SCCs*)
- 5:     **if**  $mdc > Community\_Density$  **then**
- 6:       **for**  $SCC_i \in Community$  **do**
- 7:          $Community\_Results \leftarrow$  COMPONENTVERTICESFORMNEWCOMMUNITY( $SCC_i$ )
- 8:       **end for**
- 9:     **else**
- 10:       *do\_nothing*
- 11:     **end if**
- 12:   **else**
- 13:     process *next\_community*
- 14:   **end if**
- 15: **end while**

---

### 2.3.4 Python Algorithms

#### RDyn

Dynamic networks can be used to model a wide range of real-life phenomena. However, being able to access dynamic datasets having ground-truth communities is not trivial. A classic way to address the lack of coherently annotated dataset is to employ synthetic benchmarks, such as LFR(Lancichinetti et al., 2008). For the dynamic scenario, however, only a few network generators with planted (and evolving) community structure have been proposed so far. Among them, we recognize RDyn (Rossetti, 2017). RDyn is designed to allow its user to deeply



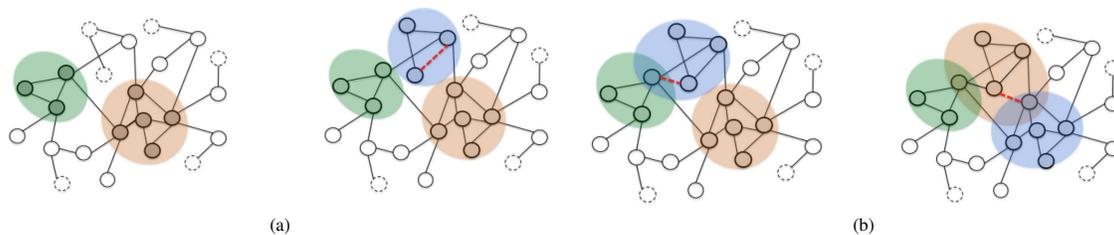
(a)

**Figure 7.** RDyn execution timeline: ground-truth communities are generated only during stable iterations (black circles). Interactions between two consecutive stable iterations compose a snapshot (here identified with  $S_0, \dots, S_5$ ). Interaction dynamics (as well as community ones) happens between consecutive iteration. Due to their definitions stable iterations are not bounded to appear with fixed displacement.

customize the generated topology and related dynamics: it was designed to generate dynamic graphs respecting to the following characteristics:

- (i) power law vertex degree distribution;
- (ii) power law community size distribution;
- (iii) tunable community quality (i.e. minimum conductance, high modularity, high density...);
- (iv) edge appearance/vanishing handling;
- (v) user-defined distribution for interaction decay and
- (vi) communities merge/split dynamics.

As shown in Fig. 7, RDyn is implemented as an iterative process since the topologies it generates are the results of subsequent choices made by the vertices within it: more specifically, during every iteration the network vertices are enabled to perform a specified set of actions (i.e., create/destroy edges—all subject to specific rules). Moreover, once completed each iteration the status of the resulting communities is evaluated, returned if considered stable, and community dynamics are planted.



**Figure 8.** Example of Tiles community growth. Each new interaction is depicted with a red dashed line. Colored shapes identify core communities. Vertices with solid borders outside the colored shapes are the peripheral vertices. Vertices with dashed borders outside the colored shapes are not involved in any community.

## Tiles

Social interactions determine how communities form and evolve. Indeed, the rising and vanishing of interactions can change the communities' equilibrium. A common approach in literature (Rossetti and Cazabet, 2018) to address topology dynamics is to:

- (i) split the network into temporal snapshots;
- (ii) repeat a static community detection for each snapshot and;
- (iii) study the variation of the results as time goes by.

This approach introduces an obvious issue: which temporal threshold has to be chosen to partition the network? This problem, which is context dependent, also adds another one: once the algorithm is performed on each snapshot how can we identify the same community in consecutive time slots? To overcome these issues, Tiles was introduced in (Rossetti et al., 2017): a Dynamic Community Discovery algorithm that does not impose fixed temporal thresholds for the partition of the network and the extraction of communities.

Tiles analyzes an interaction stream and, every time a new interaction is produced by a given streaming source, it applies a label propagation procedure to diffuse the changes to the vertex surroundings and adjust the neighbors' community memberships (Figure 8). A vertex can belong to a community with two different levels of involvement: peripheral membership and core membership. Only core vertices are allowed during the label propagation phase to spread community membership to their neighbors.

### 2.3.5 Cpp Algorithms

Algorithm 1 presents the pseudo-code of the proposed dynamic community detection algorithm based on Louvain algorithm. The algorithm input parameters are the initial network ( $G = (V, E)$ ) and the list of edges to be added and removed from the graph during the iterations ( $A$  and  $R$  respectively). For storing the community information, the algorithms use two internal community networks: the lower-level network ( $C_{ll}$ ) where the original network is maintained, and the upper-level network ( $C_{ul}$ ) where the aggregated community network is stored. The main algorithm procedure (Line 4) handles the main flow and the several subprocedures for specific algorithm tasks. These tasks subprocedures are separated into two types. The subprocedures type that do not change the network and are used to get data from both the lower-level and upper-level network (eg.: `CommunityChangedVertices()`, Line 11), and subprocedures that update the networks in terms of edges or vertices and/or community assignment (eg.: `DisbandCommunities()`, Line 19 or Line 26). The task procedures of the algorithm are conceptual and aggregate subprocedures to complete specific tasks (i.e., Adding edge to the  $C_{ll}$ ). They are repeated until a modularity increase is possible or edges to be removed or added.

**Procedure P1a:** Adding edge to  $C_{ll}$ , consists in the retrieval of a list of affected vertices and respective communities by the addition of an edge (Line 17) and by the addition of the edge itself to  $C_{ll}$  (Line 18);

**Procedure P1b:** Removal of edge in the  $C_{ll}$ . This procedure consists in the retrieval of a list of affected vertices and respective communities when the removal of an edge is performed (Line 24), and by the removal of the edge itself to  $C_{ll}$  (Line 25);

**Procedure P2:** Disband Affected Communities in  $C_{ll}$ . Based on the list of affected vertices and respective communities retrieved by `AffectedByAddition()` or `AffectedByRemoval()` the affected communities will be disbanded in  $C_{ll}$  (Line 19 or Line 26);

**Procedure P3:** Update the  $C_{ul}$  with changes of  $C_{ll}$ . The list of affected vertices and respective communities retrieved by `AffectedByAddition()` or `AffectedByRemoval()` will be also used to replicate the changes in community structure to the  $C_{ul}$  (Line 20 or Line 27). Notice that in this procedure, the added or removed edges will also be updated in the  $C_{ul}$ ;

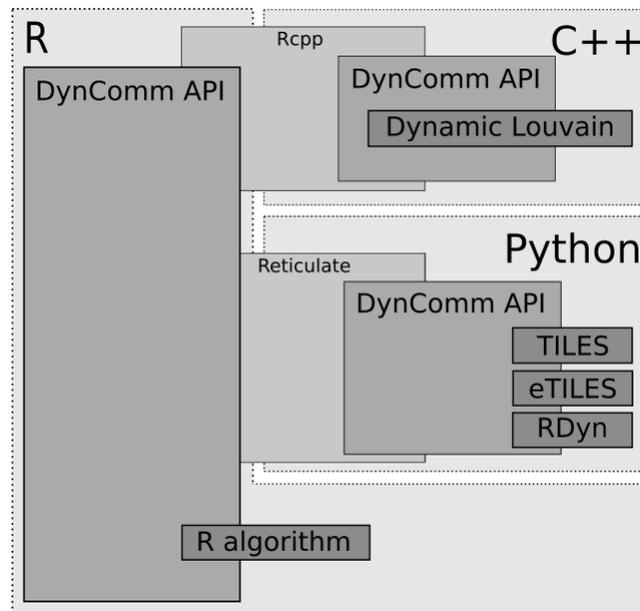
**Procedure P4:**  $C_{ul}$  will be used to perform the Louvain Algorithm Step 1 and calculate the changes in the community structure that may lead to a locally optimised modularity (Line 10);

**Procedure P5:** Update  $C_{ll}$  with the communities that changed by applying the Louvain Algorithm Step 1 to  $C_{ul}$  (Line 12);

**Procedure P6:** Use the  $C_{ul}$  to perform the Louvain Algorithm Step 2 and aggregate communities (Line 14).

## 3 DynComm R Package

DynComm package, although an R package, has interface with other languages. “Rcpp” package was used to interface C++ source code with R (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2017). To make the interface with the Python language, we used the “reticulate” package (Allaire et al., 2017). To take measurements of processing and memory use we used package “microbenchmark” (Mersmann, 2018). Figure 9 presents a block diagram of the internal setup of DynComm.



**Figure 9.** DynComm R package Architecture

### 3.1 R

R is the main language of the package, used to develop the interface, and the bridges to other languages. Since this package is related to graph analysis, some use of particular and specific packages is expected. To compute graph-based operations in R, for example, we used the “igraph” package (Csardi and Nepusz, 2006). In the following subsections, we will deal with the R user interface and an algorithm, that although not a community detection algorithm, might be used independently after the use of any community detection algorithm available, after detection of communities.

#### 3.1.1 User Interface

This package tries to supply a unified, minimalist interface that is flexible enough to allow to pass any type of information to the algorithms and retrieve the results. This holds for both end users and developers of algorithms. Since the main target of this package are community detection algorithms capable of working in a stream mode, the interface for the end user is:

- a constructor that instantiates a DynComm object to hold the choice of algorithm, any required parameters, and data
- an addRemoveEdges function to interactively add or remove edges
- a function to get the resulting vertex to community mapping
- a function to get the quality of the current mapping
- and a function to get the total incremental time taken in processing

Actually, since the graph can be very big and could be cumbersome to display the entire result at a time, there are a number of auxiliary functions that allow to get the result in smaller pieces of information, like getting the communities and then getting the vertices for each individual community, one community at a time.

### 3.2 Other Languages

DynComm R package has some examples of other languages use. Included in the package, at the present version of release, we have three Python algorithms, and one C++ algorithm. Please check the following subsections for more information.

#### 3.2.1 Programming API

For algorithm developers, their implemented code must only mimic the end user interface. It must:

- add their algorithm to the list of algorithms
- add the required parameters to the list of parameters with default values
- instantiate an appropriate object in the constructor of the DynComm object, where it will receive a reference to the graph along with any parameters required that were passed by the user, or default values if they were not.
- implement a function, or functions, to receive edges to add/remove and insert those functions inside the addRemoveEdges function
- optionally, implement functions to get the results, if they require extra processing before displaying. This is stated as optional because, usually, the results are taken directly from the graph and it is stored externally to the algorithm. Any additional processing must not change the result.

To help developers of algorithms implemented in languages other than R, both C++ and Python have an interface implemented, in the respective languages, that mimics the interface in R. This facilitates implementation because the developer needs to fiddle with neither Rcpp nor Python, which respectively perform the translation of C++ and Python to R.

## 4 Conclusions and Future Work

This document is a publication concerning the launch of the DynComm R package. It is a package developed for everyone interested in using R to do community detection in dynamic networks. The package is prepared to be improved in the future, with new algorithms added to the already interesting menu of available algorithms.

Soon, we expect to add at least a Java language algorithm. All this package code will be available in GITHUB repository <sup>2</sup>, in development and public mode to all developers. We are expecting contributions from anyone interested in promoting their algorithm(s).

## Acknowledgments

This work was fully financed by the Faculty of Engineering of the Porto University. Rui Portocarrero Sarmento also gratefully acknowledges funding from FCT (Portuguese Foundation for Science and Technology) through a Ph.D. grant (SFRH/BD/119108/2016).

## References

- Aggarwal, C. and Subbian, K. (2014). Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR)*, 47(1):1–36.
- Allaire, J., Ushey, K., Tang, Y., and Eddelbuettel, D. (2017). *reticulate: R Interface to Python*.
- Babcock, B., Babu, S., Datar, M., Motwani, R., and Widom, J. (2002a). Models and Issues in Data Stream Systems. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 1–16, New York, NY, USA. ACM.
- Babcock, B., Datar, M., and Motwani, R. (2002b). Sampling from a Moving Window over Streaming Data. In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '02, pages 633–634, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).
- Cordeiro, M., Sarmento, R. P., Brazdil, P., and Gama, J. (2018). Evolving networks and social network analysis methods and techniques. In *Social Media and Journalism-Trends, Connections, Implications*. IntechOpen.
- Cordeiro, M., Sarmento, R. P., and Gama, J. (2016). Dynamic community detection in evolving networks using locality modularity optimization. *Social Network Analysis and Mining*, 6(1):1–20.

<sup>2</sup>Available Code at <https://github.com/softskillsgroup/DynComm-R-package>.

- Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*:1695.
- Datar, M., Gionis, A., Indyk, P., and Motwani, R. (2002). Maintaining stream statistics over sliding windows. *Proceedings of the thirteenth annual {ACM-SIAM} symposium on Discrete algorithms*, pages 635–644.
- Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7.
- Eddelbuettel, D. and Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, 5:e3188v1.
- Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.
- Gama, J. (2010). *Knowledge Discovery from Data Streams*. Chapman & Hall/CRC, 1st edition.
- Gehrke, J., Korn, F., and Srivastava, D. (2001). On computing correlated aggregates over continual data streams. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data - SIGMOD '01*, pages 13–24.
- Hollocou, A., Maudet, J., Bonald, T., and Lelarge, M. (2017). A linear streaming algorithm for community detection in very large networks. *CoRR*, abs/1703.02955.
- Kim, H. and Anderson, R. (2012). Temporal node centrality in complex networks. *Physical Review E*, 85(2):026107.
- Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110.
- Leung, I. X. Y., Hui, P., Liò, P., and Crowcroft, J. (2009). Towards real-time community detection in large networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 79(6):1–10.
- Mersmann, O. (2018). *microbenchmark: Accurate Timing Functions*. R package version 1.4-6.
- Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106.
- Rossetti, G. (2017). Rdyn: graph benchmark handling community dynamics. *Journal of Complex Networks*, 5(6):893–912.
- Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks: a survey. *ACM Computing Surveys (CSUR)*, 51(2):35.
- Rossetti, G., Pappalardo, L., Pedreschi, D., and Giannotti, F. (2017). Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning*, 106(8):1213–1241.
- Sariyüce, A. E., Gedik, B., Jacques-Silva, G., Wu, K., and Çatalyürek, Ü. V. (2016). SONIC: streaming overlapping community detection. *Data Min. Knowl. Discov.*, 30(4):819–847.
- Sarmiento, R. P. (2019). Density-based Community Detection/Optimization. *arXiv*.
- Wang, C.-D., Lai, J.-H., and Yu, P. S. (2013). Dynamic Community Detection in Weighted Graph Streams. *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 151–161.
- Yun, S., Lelarge, M., and Proutière, A. (2014). Streaming, memory limited algorithms for community detection. *CoRR*, abs/1411.1279.
- Yun, S. and Proutière, A. (2014). Community detection via random and adaptive sampling. *CoRR*, abs/1402.3072.