Community identity in a temporal network: A taxonomy proposal

Luis R. Pereira[a], Rui J. Lopes[b] and Jorge Louçã[a]

[a] *ISTAR, Iscte - Instituto Universitário de Lisboa, Avenida das Forças Armadas, Lisboa 1649-026, Portugal*
[b] *Iscte - Instituto Universitário de Lisboa, Avenida das Forças Armadas, Lisboa 1649-026, Portugal*
Instituto de Telecomunicações, Torre Norte - Piso 10, Avenida Rovisco Pais 1, Lisboa 1049-001, Portugal

ARTICLE INFO

Keywords:
Complex systems
Networks
Clustering

ABSTRACT

Communities of nodes are one of the most important meso structures in a network. In static networks they are essentially characterized by the nodes they hold, how modular they are and how they connect to other communities. Once identified, they do not change. In networks that evolve over time, communities can shed and acquire new nodes. This generates new constructs and raises the question of community identity, and of the characterization of the events that define their lifecycle. Although researchers have devoted efforts to address some of these questions, we believe that a formalized classification and a principled method to identify community events is still lacking. In this paper we propose such a classification in the form of a robust taxonomy, supported by a similarity metric based on the Jaccard index but adjusted to chance, and a set of rules that unequivocally can track a community journey from "cradle to grave".

1. Introduction

Identity preservation is a general problem of any complex, time evolving system. The Ship of Theseus paradox is one of its most famous illustrations, arching back to the Greek mythology. Theseus, an Athenian hero, returns to Athens in glory after defeating the Minotaur on the island of Crete. In his honor, the ship in which he sailed is kept in a museum, where, due to the ravages of time, its original parts are substituted as they rot. Eventually all end up being replaced. Can we still consider that this is the same ship Theseus sailed on? If not, when did it stop being so? This thought experiment has been discussed by many philosophers, spanning millennia, from Heraclitus to John Locke [13].

If we consider what constitutes a network community and how it evolves in a temporal network, we are faced with a similar problem. Can a community that shares no nodes with one previously observed, be the same community? If not, and assuming granular step changes, it must have lost its identity at some stage. A fundamental issue thus becomes what criteria to use to make that determination.

To clarify, here we are not talking about absolute identity, or what Leibniz called "The Identity of Indiscernibles" [20]. That is, x and y are identical, if and only if for every attribute A, its existence in x implies its existence in y, or formally $\forall A(A \in x \leftrightarrow A \in y) \rightarrow x = y$. Under this definition, those ships are absolutely discernible. We are really talking about relative identity, the same that allows us to identify an adult as a child of yore, or a soccer club as the same club with a totally different roster of players and technical staff years later. This may appear as a simple semantic question, but it is in fact an important distinction, especially when it comes to two aspects of communities in temporal networks: their detection and their identification. In this article we are especially concerned with the latter, and how it relates to a lifecycle of events that group together a set of detected communities under the same (relative) identity.

In static networks the identity of a community can be described as a surjection from the node set to the community set, an onto mapping establishing the node community membership. As we extend our study of networks exhibiting community structure into the temporal domain, communities are no longer static. A community that is observed at a given moment may be different later on. Representing the ground truth of such a network as a timed-sequence of surjections may faithfully represent the community structure overtime, but does not lead unequivocally to the understanding of its lifecycle. For that we need an accepted taxonomy of lifecycle events, and methods to correlate the changes in the community structure to those events. In general, classifying events is not a solved problem and formalization is lacking. Furthermore, recovering lifecycle events may not be totally possible without information not inherently present in the network topology, which precludes a non-parametric solution to a problem where network topology is the only available data. In this article, we present a formalized taxonomy and propose a method to track community evolution assisted by meta information.

Communities are a challenging network construct. Although they are commonly defined as a set of nodes that are more densely connected among themselves than to the rest of the network, the fact is that, given a network, determining if and how many communities exist in that network may not have a single, clear answer [11]. Temporal networks usher in an additional layer of complexity, which, nevertheless, has not deterred many authors from trying.

Let us note that, here, we are not directly concerned about what constitutes a community, but how it evolves in time. In this context, a clustered network is just a set of sets.

Expanding the ground truth of community structure to include events of a temporal nature is not a new topic. Barabási in his book Network Science [5] summarizes current con-

*Corresponding author
ramada.pereira@isce-iul.pt (L.R. Pereira)
ORCID(s): 0000-0001-9115-7959 (L.R. Pereira)

sensus on what these events should be. It documents six elementary events: Growth, Contraction, Merging, Splitting, Birth and Death.

We believe however that this consensus is problematic. For instance when defining a community split, where do you draw the line between a split and a contraction? Is losing one node, a split? If not, how many? And how would one classify an event of a community that fully fragments, shedding nodes to multiple communities, which in turn receive nodes from several other communities? In our work we came to believe that topology alone cannot answer these questions. Depending on subject domain, a community may cease to exist as a separate entity when none of its nodes are seen after a given time T or when a given fraction of its members disappear. Here, the network topology does not shed any light. Examples from the real world abound, consider the minimum quorum for a shareholder assembly or the level of infestation by an indicator species in biology. In both of these cases, external information is required to validate the existence of a functioning community.

We also find that it is easier to reason about community events anchored on the community and not on the event. So, for example, an event where many nodes change community membership, may result in a community fragmenting while other communities in the same network may grow by acquiring some of its fragments.

In support of this approach we define three simple top level community events: Birth, Continuation and Death. That is, once born, a community either continues or dies. Continuation will have different meanings depending on context. In an abstract way, however, we define it as *similarity beyond a cutoff point* allowing recognition of a former community in a present one. We propose a similarity metric based on the Jaccard index [17] to compare communities, with a parametric cutoff point dependent on subject domain. If the metric, as a distance function, between any two communities taken from community sets at t and $t+\epsilon$ clears the threshold, then the oldest continues in the most recent. Note that a community may continue in multiple communities depending on their similarity. That multiplicity together with time orientation further classifies the continuation event. For example: given two communities at time t, c_i^t and c_j^t, and two communities at time $t+\epsilon$, $c_k^{t+\epsilon}$ and $c_l^{t+\epsilon}$, c_i^t can continue in $c_k^{t+\epsilon}$ and $c_l^{t+\epsilon}$ (a split), while $c_l^{t+\epsilon}$ is a continuation of c_i^t and c_j^t (a merge). This simplifies the model, catering for the complexity of the multiple types of events that can occur in the clustering of a temporal network, defining events from a community point of view, allowing for domain specific external input that further characterizes the community lifecycle.

In the reminder of this article we refer and review related work that predates our current proposal in section 2. In section 3 we describe how we compare communities to determine lifecycle events and their taxonomy tree. In section 4 we introduce the adjusted Jaccard index, the metric we used to compute the distance between pairs of communities, and the null model that supports it. In section 5 we present the full classification methodology and procedure, followed in section 6 by examples using a toy model and an empiric network. We conclude in section 7 with future directions and follow-on work.

Throughout the text we use a consistent notation, using **C** to identify a community as a set of nodes, and **S** to identify a multiset of community cardinalities. Both can optionally be superscripted to specify a given network observation. If denoted in lowercase, they refer to a single community that can additionally be subscripted for identification. The usage of upper and lower case is consistently used to differentiate a collection from its elements. Other notation will be introduced as required.

2. Related Work

In spite of their obvious applicability in representing time evolving complex systems, temporal networks studies are still under represented in the overall complex network scientific production. The subtopic of communities in networks is no exception, even though in the last decade or so, a number of proposals have been put forward to define and detect what a community is, in the context of an evolving network.

A simple example of community detection in a temporal network can be found in [18], where authors add inter-time edges to the network, connecting the same and related nodes at successive moments, followed by traditional static community detection on the resulting aggregated network. This results in a partition of the network that may identify enduring communities, but is of limited use when examining a particular observation of the network or to understand how a community evolves.

Static community detection is usually performed by optimizing a quality or fitness score, such as modularity, conductance, size of compressed information flows, among many others. Unless the community is frozen in time, changes will affect that score. Many authors extend the fitness score to smooth community evolution[3], usually by establishing additional objectives, such as minimizing the clustering changes across time thru measures such as the normalized mutual information [9], or by including past, and sometimes future, network observations in the fitness scoring function. This smoothing has the additional advantage of mitigating algorithmic artifacts, as most fitness functions are frequently computationally complex to optimize, usually through heuristics that are sensitive to initial conditions and computing effort.

In a temporal network, approaches to community detection usually follow one of two options: they either consider each network observation independently or directly combine multiple observations. The way this is performed varies and authors in [2] distinguish between:

- two-stage approaches, where detection is performed per observation complemented by partition matching with previously identified communities;
- evolutionary clustering, where detection over the current observation is a function of the observed topology and of prior community structure, usually optimizing

a modified quality function that dampens the influence of previous observations as they fade in time;

- and methods that couple all observations into a single network, usually by linking nodes across observations, and perform community detection on the consolidated network.

In their survey [28] authors expand on this classification, creating a hierarchy of approaches, that at the first level is similar to the one in [2], defining, respectively "Instant Optimal", "Temporal trade-off" and "Cross time" approaches but providing additional granularity by detailing subcategories for each class. A full survey is beyond the scope or intent of this article. The reader is referred to [4, 29, 33, 15, 10, 28, 8] for more information.

Although most of these efforts concentrate on identifying temporal communities in an absolute sense, in this article and this section we are especially concerned with relative identity and on how communities evolve from birth to death. From this standpoint, and in the strict context of our taxonomy proposal, the way a community is identified is immaterial. Our proposed approach works regardless. This does not imply that network evolution cannot contribute to community detection, as many authors have proposed, resulting in methods and algorithms that simultaneously try to detect communities and classify the events they endure. We have not found, however, any article that exclusively focus on lifecycle analysis.

To our knowledge, community events were first proposed in [23] and, since then, there seems to be an emergent consensus around events like birth, merge, split, growth, expansion, contraction and death. Some authors propose additional events like continuation (i.e. no growth or expansion) and resurgence for communities that appear periodically [28]. A summary of these events with informal definitions can be found in [7] as well as a formalism for lifecycle representation based on a directed graph where nodes are timed community observations and edges are continuation events bridging time gaps.

When matching communities for event determination many authors use, as we do, a set based distance measure. The Jaccard index [17]:

$$J(c_i^t, c_j^{t+\epsilon}) = \frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_i^t \cup c_j^{t+\epsilon}|} \tag{1}$$

is used by authors in [14, 21, 22, 23], even though it may be named differently in some cases. In [31] authors use different measures depending on event, such as the ratio of the size of the intersection to the size of the largest community, basically a measure of dilution, to determine whether a community is born or vanished, or the relative size of the proper subset of a community in a subsequent timestep to determine continuation. In [32] the same authors distinguish between communities and metacommunities, the latter being a construct to track community evolution. In [1], continuation is predicated on set equality of community membership at succeeding time steps, while merge and split depend on dilution of nodes gated by individual community contribution for the event. The appearance of new communities (which the authors name "Form") and disappearance ("Dissolve") are conditioned on, respectively, no prior or post observation of any of the nodes on the formed or dissolved community. In [16] authors use a measure that favors communities similarly sized with a high ratio of common nodes:

$$similarity(c_i^t, c_j^{t+\epsilon}) = min\left(\frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_i^t|}, \frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_j^{t+\epsilon}|}\right) \tag{2}$$

A different approach is taken in [6] where a method (*GED*) is proposed where the measure used is the forward dilution of a community ($\frac{|c^t \cap c^{t+\epsilon}|}{|c^t|}$) modulated by the relative "social position" of surviving member nodes, basically non topological information assigned to specific nodes, changing their relative weight in community formation. An approach based on forward and reverse dilutions, but without any additional adjustments, can be found in [30], where the results of applying the dilution formulas to all community pairs at succeeding network observations are used to build correlation matrices, that are then subject to a parametric process to determine lifecycle events. In [19], authors classify lifecycle events using a directed weighted network where nodes are observed communities and edges connect related communities, weighted by the fraction of surviving nodes as communities evolve. With the exception of [21] (which we analyse further in section 6.2), all of the prior approaches, including our own, identify lifecycle events depending on user specified parameters. In fact, we believe that the definition of a community event, with exception of clear cut cases, such as, for example, when a totally new and cohesive set of nodes appear on the network as a birth event, requires meta information not inherent in the network topology.

Our approach is not dissimilar to the one adopted in [14], but with a distance measure adjusted for chance. We also simplify the concept of community evolution, by anchoring it on the community itself at a given point in time and not on the network. Like most other approaches, ours is parametric, requiring meta information about community relative identity.

3. Recovering Community Events

Clearly defining community events is useful for many reasons, such as the development and testing of dependable temporal community detection and evolution algorithms.

Our lifecycle identification framework addresses the problems associated with the classification of complex events when nodes exit and enter various communities as well as comprehensively covering other events relevant in the various problem domains where temporal networks play a role.

On this basis, we created a hierarchical, multi-level classification scheme, based on the following rules:

- Once born into existence, a community either continues or dies.

- A community continues in another community if their measured similarity clears an externally supplied cut-off. A consequence of this rule is that remains of a community that do not reach the threshold for continuation, contribute to newly born communities or the expansion of others or both.

- Single continuation events, that is, continuation events that involve only a pair of communities, can be further subdivided into:

  - Growth and contraction events with net acquisition or loss of nodes.
  - Replace events when the communities keep the same cardinality, but with some of their nodes replaced.
  - Preserve events if no changes in node membership occur.

- Multiple continuation events, that is, continuation events that involve more than a pair of communities, can be further subdivided into:

  - A split, if a multiple continuation event is observed from the past.
  - A merge, if a multiple continuation event is observed from the future.

- A community can die either if its nodes are no longer seen on the network (vanishes), or it does not continue in any other community (absorved). A community can experience loss of nodes and absorption simultaneously and the proper classification would, in our proposal, follow the largest of the remaining and dead node set sizes.

- A community can be born from new nodes (beginning) or from fragments of other communities (regenerated). Both can happen simultaneously and classification follows the largest node set.

- A community can also resurge on the network, for example on cyclic events. This is detected as a single continuation bridging a lapse of time longer than the network temporal resolution and can potentially occur on "Begin", "Regenerate", "Grow", "Contract", "Replace", "Split" and "Merge" events.

A full taxonomic tree is depicted in figure 1. The method for community continuation analysis as presented in the next section abides by the above categorization.

To compare community similarity many authors use the Jaccard index (J) [17], as previously mentioned. Authors in [23], call it the auto-correlation function and extend it to any time delta. J varies from 0, when no nodes are common between communities, to 1 when all nodes are shared. Intuitively, it expresses similarity between sets. However, in a potentially constrained domain, such as in a temporal network where nodes persist across time, its interpretation should be subject to probabilistic scrutiny. For this reason, we propose the usage of an adjusted Jaccard index ($\hat{J}$) to compare communities, as described in the following section.

4. Adjusted Jaccard index and null model

A random network should not exhibit community structure[1]. Similarly, a random redistribution of community membership across the node set over $t \rightarrow t + \epsilon$, should result in a null similarity index between any pair of communities $\in C^t \times C^{t+\epsilon}$. However, this redistribution will result in an average positive J of all community pairs in anything but the asymptotic limit of network size. To correct for this, we introduce an adjusted Jaccard index ($\hat{J}$). To compute $\hat{J}$, we make use of auxiliary "null version" variables which we denote with a ˘ accent.

Given two multisets S^t, $S^{t+\epsilon}$, with $\sum S^t = \sum S^{t+\epsilon}$, a random assignment of nodes $V \mapsto \check{C}^{t+\epsilon}$, subject to[2]:

$$\mathbb{P}(v \in \check{c}_i^{t+\epsilon}) = \frac{s_i^{t+\epsilon}}{\sum S^{t+\epsilon}} \tag{3}$$

results in an expected number of shared nodes between pairs $\in C^t \times \check{C}^{t+\epsilon}$:

$$\mathbb{E}(|c_i^t \cap \check{c}_j^{t+\epsilon}|) = s_i^t \times \frac{s_j^{t+\epsilon}}{\sum S^{t+\epsilon}} \tag{4}$$

for any community c^t, and the corresponding community $\check{c}^{t+\epsilon}$ built from the probabilistic distribution of nodes onto $\check{C}^{t+\epsilon}$ resulting from equation 3. Let's notate this $f_\emptyset(c^t, \check{c}^{t+\epsilon})$, as we will use it to develop the adjusted Jaccard index.

Consider two communities $c_i^t, c_j^{t+\epsilon}$. We propose a null model to adjust their J in such a way that,

1. $|c_i^t \cap c_j^{t+\epsilon}| \leq f_\emptyset(c_i^t, \check{c}_j^{t+\epsilon}) \Leftrightarrow \hat{J} = 0$
2. $c_i^t \subseteq c_j^{t+\epsilon} \vee c_i^{t+\epsilon} \subseteq c_j^t \Leftrightarrow \hat{J} = J$

The first adjustment captures the intuition that a random distribution of nodes should not lead to affinity between communities. The second adjustment captures the intuition that the index should not be adjusted if the community is preserved, or if its nodes are kept together or isolated from the rest of the network. A consequence of these adjustments is that $\hat{J} \in [0, J]$.

To implement the first adjustment we compute the Jaccard index between communities $c_i^t, \check{c}_j^{t+\epsilon}$, under the conditions of equation 4, basically the ratios of the intersection with the union of communities c_i^t and $\check{c}_j^{t+\epsilon}$. We denote this index as $\check{J}$:

$$\check{J}(c_i^t, \check{c}_j^{t+\epsilon}) = \frac{s_j^{t+\epsilon} \times s_i^t}{\sum S^{t+\epsilon} \times (s_j^{t+\epsilon} + s_i^t) - s_j^{t+\epsilon} \times s_i^t} \tag{5}$$

[1] This fact is the basis of one of the most popular methods of community detection [12]

[2] Nodes trivially appear and vanish in many temporal networks, resulting in a variable number of nodes as time evolves. When that happens in two consecutive observations at $t, t+\epsilon$, we add an additional fictitious community of new born nodes at time t and another community of dead nodes at time $t+\epsilon$ thus avoiding handling network samples of different cardinality.

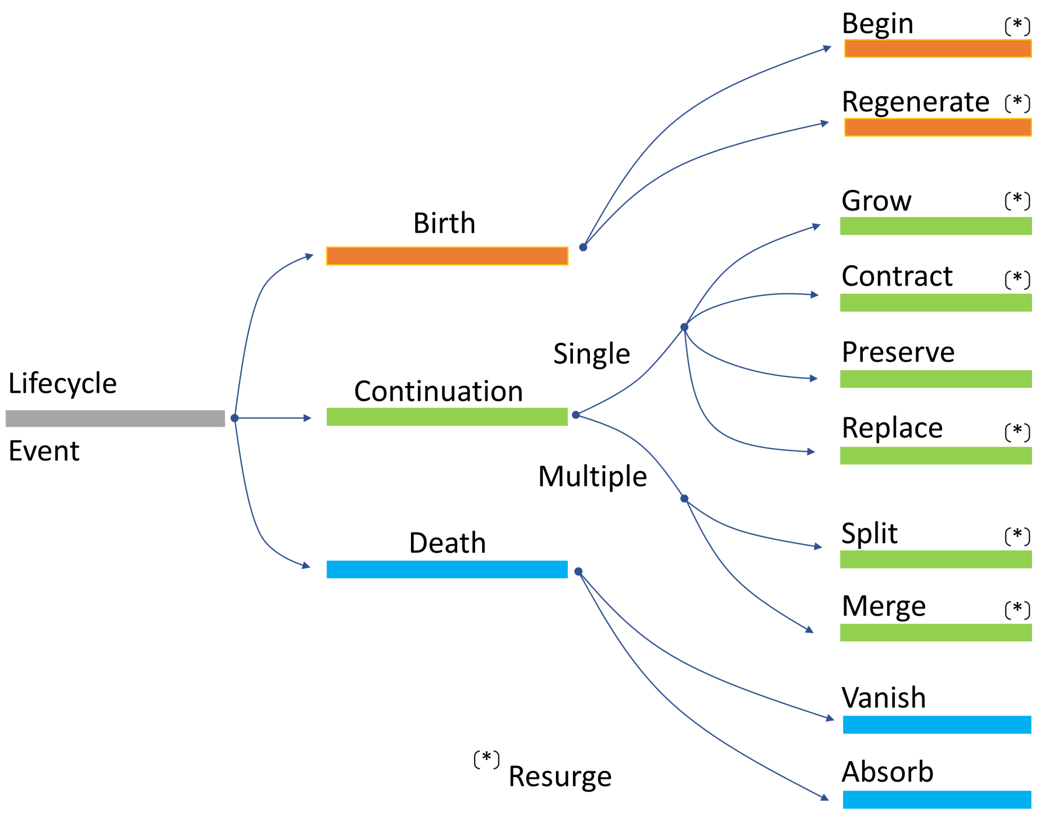


Figure 1: Events in the lifecycle of a community in a temporal network. Classification dependent on multiplicity of continuation events and relative set sizes

Formula 6 allows us to correct the index to zero on random chance, while preserving a perfect score of "1" when $c_i^t = c_j^{t+\epsilon}$:

$$\max\left(\frac{J(c_i^t, c_j^{t+\epsilon}) - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon})}{1 - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon})}, 0\right) \tag{6}$$

However, this will adjust down the index when c_i^t is a proper subset of $c_j^{t+\epsilon}$ or vice-versa, contrary to our null model design. To enforce our model, we compute the Hadamard product ($\breve{J} \odot R$) where R is the "proper subset coefficient" matrix, with elements defined as:

$$r_{ij} = 1 - \frac{|c_i^t \cap c_j^{t+\epsilon}| - f_\emptyset(c_i^t, \breve{c}_j^{t+\epsilon})}{\min(s_i^t, s_j^{t+\epsilon}) - f_\emptyset(c_i^t, \breve{c}_j^{t+\epsilon})} \tag{7}$$

$r_{ij} = 0$ if a proper subset condition exists, increasing $\propto (\min(s_i^t, s_j^{t+\epsilon}) - |c_i^t \cap c_j^{t+\epsilon}|)$.

The proposed adjusted index now becomes:

$$\hat{J}(c_i^t, c_j^{t+\epsilon}) = \frac{J(c_i^t, c_j^{t+\epsilon}) - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon}) \times R(c_i^t, c_j^{t+\epsilon})}{1 - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon}) \times R(c_i^t, c_j^{t+\epsilon})} \tag{8}$$

We studied empirically the behaviour of our adjusted Jaccard index. From our previous discussion, a random distribution of nodes by communities, should, in principle, result in a null similarity score between any pairs of communities from two succeeding network observations. If we were to plot the cumulative distribution function (*cdf*) of the average similarity index for samples of such a network, ideally, it should result in $cdf(index) = 1 | index \in [0, \infty]$, with all observations at 0.

We tested the Jaccard index and our adjusted index on sets of random temporal network configurations, varying the number of communities and the number of nodes. The community cardialities were sampled from a powerlaw function with exponent $\gamma = 2.5$ for each observation, as this cardinality distribution is frequently observed in empiric networks, even though similar results were obtained when sampling

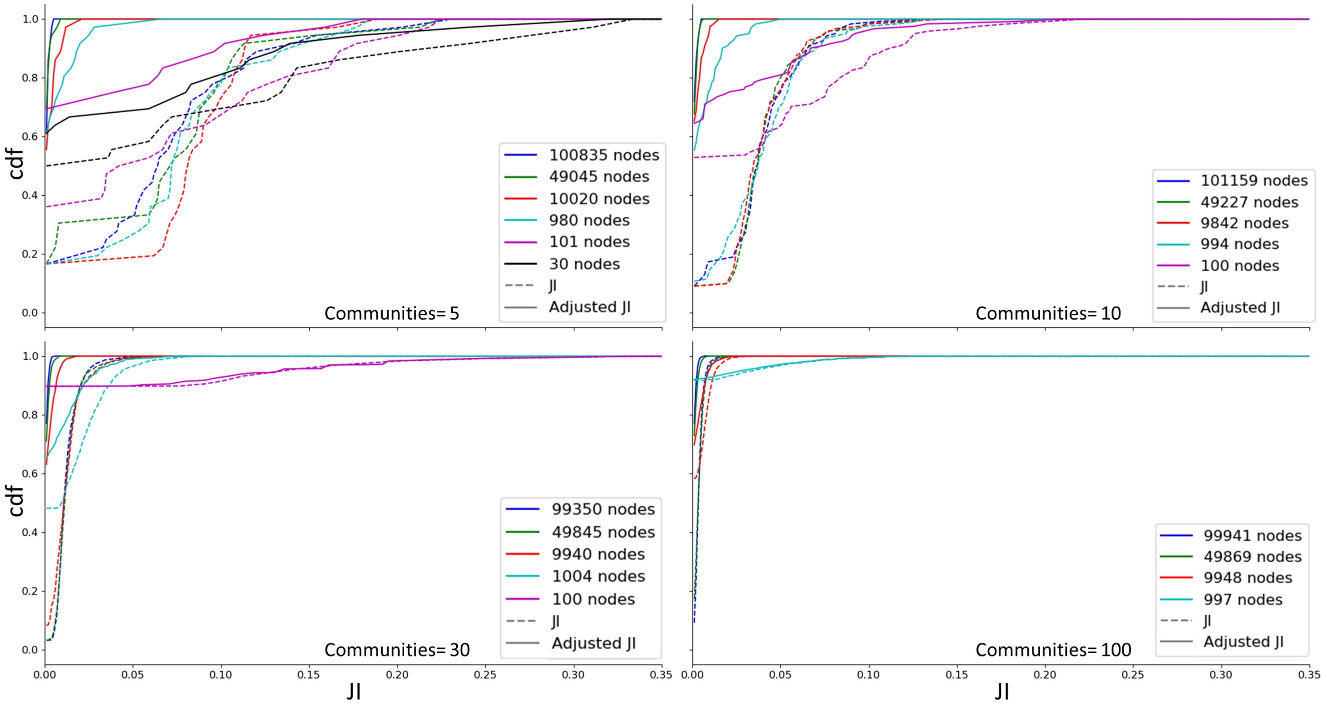


Figure 2: Performance of the Adjusted Jaccard index ($\hat{J}$) for the null model. Cumulative distribution function of $\hat{J}$ and J. Average $\hat{J}$ compared to averaged J for pairs of communities with a positive index for varying numbers of communities and nodes. Each line represents the average of 100 runs. As the number of communities increase, $\hat{J} \to J$.

from uniform distributions. 100 observations were made on each network. For each pair of observations the average of all positive indexes was computed. The resulting *cdf* of J and $\hat{J}$ can be seen in figure 2. $\hat{J}$ performs much closer to the ideal result than J when the number of communities is low. As the number of communities and nodes grow, the differences vanish and at ≈ 100 communities and ≈ 10000 nodes, there is practically no difference between the indexes and the null model ceases to be relevant, as both are close to the expected *cdf* for random transitions.

The larger divergence from the ideal behaviour on small networks is the result of two factors. With less nodes and less communities, the probability of spurious similarities increases as nodes have less degrees of freedom. That is taken care by the null model. However, discretization also plays a role as a perfect uniform distribution is not possible when moving from a continuous to a discrete domain. After all, nodes cannot be sub-divided. This explains the deviation in the *cdf* of $\hat{J}$ from the expected *cfd* on low community and node counts.

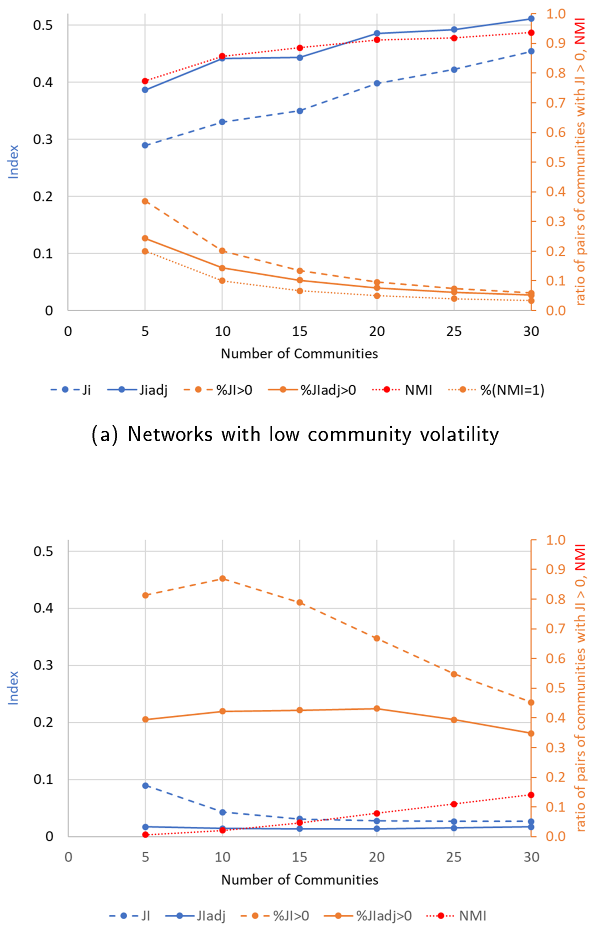
We also tested the indexes against highly stable networks, that is, networks where communities exhibit low membership turnover (see figure 3a). These networks were generated using a tool [24] that, given a network observation and a multiset of community cardinalities, flows the nodes across communities minimizing changes. Just as in the previous example, community cardinalities were sampled from a power law function with $\gamma = 2.5$. To show how close the observations were, we computed the average Normalized Mutual Information (*NMI*) across network observations. In figure 3 we plot the average positive J, $\hat{J}$, *NMI* and the average percentage of community pairs exhibiting a positive index, for 6 sets of temporal networks with 50 observations and ≈ 1000 nodes, varying from 5 to 30 communities in steps of 5. In figure 3a we also include the ratio of positive indexes for a frozen network where $NMI = 1$. For comparison, results from a randomly evolving network can be seen under the same conditions in figure 3b.

For this network size, the adjusted index reduces the number of positive scores, as a consequence of the null model application. This contributes to an increased average of positive indexes for networks with communities with low membership volatility. For random networks the model is sufficiently robust to keep a lower averaged $\hat{J}$.

5. Event categorization method

The adjusted Jaccard index ($\hat{J}$) is used in the method below to determine community continuation. We note, however, that the method is not dependent on this specific similarity measure. Others, more appropriate to a given subject domain, can be used, as long as, from the contingency matrix (see step 1 below), they produce a binary decision over community continuation.

The full method has the following steps:



(a) Networks with low community volatility

(b) Networks with random communities

Figure 3: Performance of the Adjusted Jaccard index ($\hat{J}$) for highly stable (a) and for random networks (b). Averages of 50 observations for 6 sets of networks with varying number of communities and an average of 1000 nodes. We plot positive $\hat{J}$, J, percentage of positive $\hat{J}$, J and *NMI* for each network. As can be seen in this plot, the adjusted index detects less false positives as a result of the null model adjustment. As expected, it also improves the average index, but only in the case of highly stable networks

.

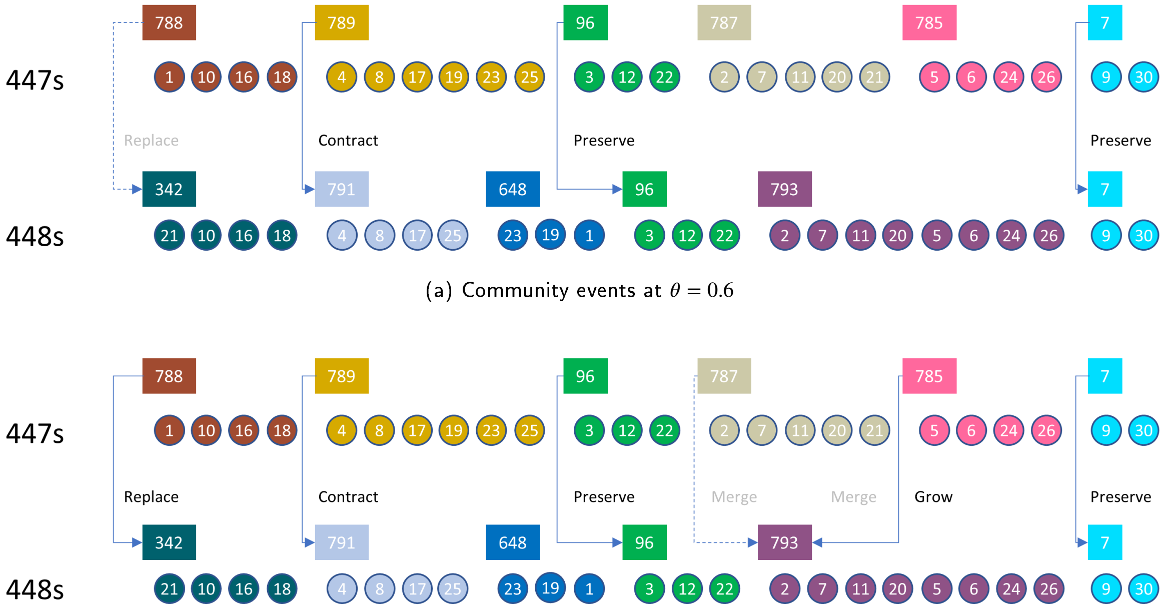
1. A confusion (or contingency) matrix X, with size $S^t \times S^{t+\epsilon}$, is created with entries $x_{ij} = c_i^t \cap c_j^{t+\epsilon}$
2. A Jaccard index matrix (J) is created from X and $S^t, S^{t+\epsilon}$ using equation 1.
3. A null Jaccard index matrix ($\breve{J}$) is created from $S^t, S^{t+\epsilon}$ using equation 5 .
4. An adjusted Jaccard index matrix $\hat{J}$ is created from J and $\breve{J}$ using equation 8.
5. An external threshold θ is applied as a cutoff binary filter over $\hat{J}$ resulting in a Boolean matrix H representing a k-adic relation between communities. We call this the continuation matrix.
6. The rows and columns of H are summed resulting in split and merge vectors P and M, respectively.
7. For every $h_{ij} = 1$ there is a single **continuation** top level event between communities c_i^t and $c_j^{t+\epsilon}$ if $m_i = p_j = 1$. This top level single **continuation** event generates second level:
   (a) **grow** event if $s_i^t < s_j^{t+\epsilon}$;
   (b) **contract** event if $s_i^t > s_j^{t+\epsilon}$;
   (c) **preserve** event if $s_i^t = s_j^{t+\epsilon} \wedge \hat{j}_{ij} = 1$;
   (d) or **replace** otherwise.
8. For every $h_{ij} = 1$ there is a multiple **continuation** top level event between communities c_i^t and $c_j^{t+\epsilon}$ if $(m_i \vee p_j) > 1$. Top level multiple **continuation** events generate second level:
   (a) **merge** events if $m_i > 1$;
   (b) **split** events if $p_j > 1$.
   Both are generated if $(m_i \wedge p_j) > 1$.
9. For every $m_i = 0$, we have a **birth** top level event for community $c_i^{t+\epsilon}$. Top level **birth** events generate second level:
   (a) **begin** events if there are more new nodes than absorbed nodes, or formally if $s_i^{t+\epsilon} \geq 2\times\sum_{j=1}^{s_j^{t+\epsilon}} x_{ij}$;
   (b) or **regenerate** events otherwise.
10. For every $p_i = 0$, we have a **death** top level event for community c_i^t. Top level **death** events generate second level:
    (a) **vanish** events if there are more dead nodes than absorved nodes, or formally if $s_i^t \geq 2\times\sum_{j=1}^{s_i^t} x_{ij}$;
    (b) or **absorve** events otherwise.
11. The events {"begin", "regenerate", "grow", "contract", "replace", "split" , "merge"} can be further classified with a **resurge** attribute as soon as a single continuation results when applying this method to older network observations in a most recent order, i.e. between pairs $(c_i^{t-n\epsilon}, c_j^{t+\epsilon})$, where n varies from 2 to $\frac{l}{\epsilon}$ where l, ϵ stand respectively for the network longevity and temporal resolution.

6. Examples

In this section we present two examples of the application of the proposed taxonomy and method. In subsection 6.1, we use a toy model to illustrate the individual steps taken to determine community lifecycle events, and, in subsection 6.2, we show some of the useful information that can be extracted by the model application to an empirical temporal network representing a soccer game, where players are nodes, and communities are sets of players in close interaction.

6.1. Toy model

To illustrate the event categorization method consider two community sets $C^t, C^{t+\epsilon}$ with 5 communities each with 20 nodes ($S = \{20^5\}$), where the flow of nodes across $t \rightarrow t + \epsilon$ is given by the following confusion matrix (step 1 of



(a) Community events at $\theta = 0.6$

(b) Community events at $\theta = 0.42$

Figure 4: Empiric network community events as determined by J and $\hat{J}$ at (a) cutoff point $\theta = 0.6$ and (b) $\theta = 0.42$. These are two observations with one second delay of all sets of players and goals on the pitch. The cutoff point can be seen as the trade off between continuations and death and birth events, and its value is subject domain dependent. The adjusted Jaccard index is more stringent on selecting continuation events as it adjusts for random chance. Dashed lines and greyed out text represent events that do not clear θ under $\hat{J}$ but do under J. Death and birth events not represented for clarity.

section 5):

$$X = \begin{bmatrix} 0 & 0 & 10 & 0 & 5 \\ 2 & 0 & 0 & 2 & 2 \\ 5 & 0 & 0 & 5 & 5 \\ 10 & 0 & 10 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 \end{bmatrix}$$

This results in a simple Jaccard matrix (step 2):

$$J = \begin{bmatrix} 0 & 0 & 0.33 & 0 & 0.14 \\ 0.053 & 0 & 0 & 0.053 & 0.053 \\ 0.14 & 0 & 0 & 0.14 & 0,14 \\ 0.33 & 0 & 0.33 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

As all communities have the same size, $f_{\emptyset}(c_i^t \cap \breve{c}_j^{t+\epsilon}) = 4$, $\breve{J}(c_i^t \cap \breve{c}_j^{t+\epsilon}) = \frac{1}{9}$ over a uniform supported random distribution of nodes across communities at time $t+\epsilon$. The adjusted Jaccard matrix then becomes (step 4):

$$\hat{J} = \begin{bmatrix} 0 & 0 & 0.30 & 0 & 0.070 \\ 0 & 0 & 0 & 0 & 0 \\ 0.070 & 0 & 0 & 0.070 & 0.070 \\ 0.30 & 0 & 0.30 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

if we take, as an example, a cutoff of $\theta = 0.2$, we get the continuation matrix (H), the split (P) and merge (M) vectors (steps 5, 6):

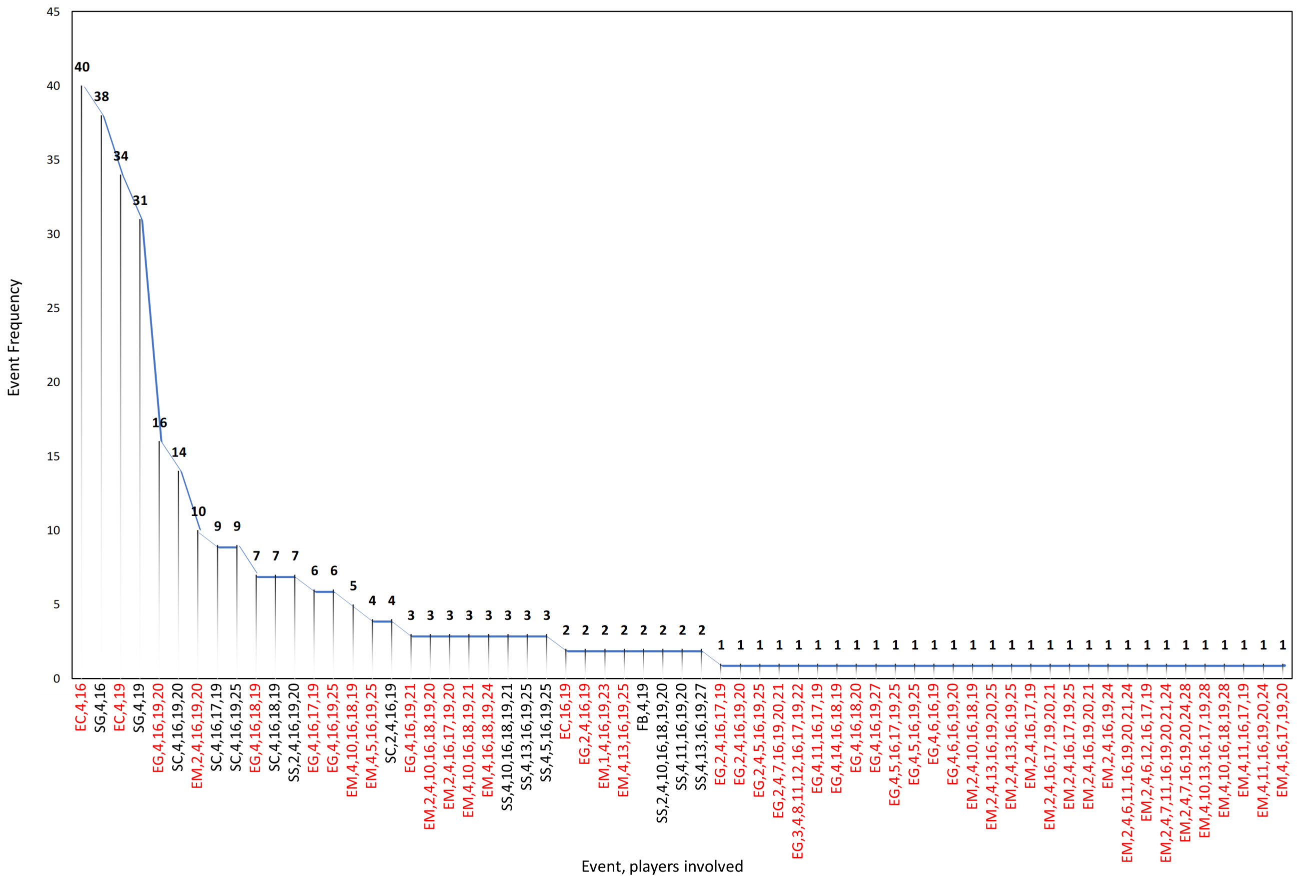
$$H = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} P = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \\ 1 \end{bmatrix}$$

$$M = \begin{bmatrix} 1 & 1 & 2 & 0 & 0 \end{bmatrix}$$

Applying steps 7, 8, 9 and 10, we have **continuation** events between $(c_1^t, c_3^{t+\epsilon}), (c_4^t, c_1^{t+\epsilon}), (c_4^t, c_3^{t+\epsilon}), (c_5^t, c_2^{t+\epsilon})$. Community c_4^t suffers a **split** and $c_3^{t+\epsilon}$, a **merge**. Communities c_2^t, c_3^t die, and communities $c_4^{t+\epsilon}, c_5^{t+\epsilon}$ are born. As $s_2^t = 20$ and $2 \times \sum_{j=1}^{5} x_{2j} = 12$, c_2^t death is a **vanish** event. As $|c_3^t| = 20$ and $2 \times \sum_{j=1}^{5} x_{3j} = 30$, community c_3^t death is a **absorption** event. Similarly, applying step 9 of the above method, we can further classify $c_4^{t+\epsilon}$ birth as a **begin** event and $c_5^{t+\epsilon}$ birth as **regenerate** event.

6.2. Application to an empiric network

The taxonomy and the event categorization method can recover information from a clustered temporal network that may not be easily apparent thru other methods. In this section, we apply it to a network resulting from sampling soccer players position on the pitch and clustering them into sets or "communities" by physical proximity. The clustering process is explained in [25]. The match is sampled at 10Hz, generating close to 60,000 observations during the whole game.



Type	Key	Event	Comments	Type	Key	Event	Comments
Exit	EA	Absortion	Does not continue in any other community	Entry	FB	Rebirth from	Reappears from an older community
	EC	Contraction to	Contracts to another community		SC	Contraction from	Continuation of a contracting community
	EG	Growth to	Grows to another community		SG	Growth from	Continuation of a growing community
	EM	Merges to	Merges into another community		SO	Replaces nodes	Rebuilds community by replacing nodes
	EO	Replaces nodes	Creates another equal sized community by replacing players		SR	Regeneration	Appears not continuing any other community
					SS	Split into	Continuation of a splitting community

Figure 5: Event frequency distribution for set 63 composed by attacking player 4, and defensive players 16,and 19 of opposing team. In this figure we can see that the most common formation and separation events occur when player 19 joins or leaves the set.

Several thousand unique sets of players are usually detected per match, but their distribution is far from uniform. Some occur quite frequently while many occur very rarely [26]. Physical proximity between players in collective ball games is a determining factor of game play and understanding how their patterns evolve can support game strategy and training [27]. This is where a community lifecycle analysis can be of value.

For this first example we use one single game transition to illustrate how the selection of the cutoff and the adjusted Jaccard index influence the categorization of events, followed by showing how a frequently detected set of players (2 fullbacks and a winger) emerges and changes.

In this dataset there is a maximum of 30 nodes (22 players in-game, 6 substitutes and 2 goals), numbered from 1 to 30, and a variable number of sets of nodes (communities) which are serially numbered as they appear in the match.

The transition from second 447 to 448 of game play is shown in figure 4. Seen thru the "lens" of our method we can see the events the sets of players underwent. At a cutoff $\theta = 0.6$ (figure 4a), set 788 is absorved using the adjusted Jaccard index, but continues when using the non-adjusted index. Although set 342 keeps $\frac{3}{4}$ of set 788 players, it still does not meet the cutoff for continuation. Conversely, set 791 does continue from 789, even though it keeps a lower ratio of players $\frac{4}{6}$. This is the result of favoring the concentration of nodes in our adjusted index.

In figure 4b we relax the cutoff point lowering it to a level ($\theta = 0.42$) where clear differences to figure 4a can be observed between lifecycle events. As expected there are more continuations. As we frequently stressed, there is nothing inherent in the network that can guide the selection of θ. It is totally dependent on subject matter expertise, in this case, how much of a compositional change a set can endure while still expressing functional affinity. Authors in [21] used a dynamic threshold that depends on the actual community structure at every timestep transition: more specifically that threshold is the minimum of the set of maximum J per community of all cross-timestep community node flows, or using our continuation matrix, it is the minimum of the maximum of the vectors P and M (step 6 of section 5). This guarantees an increase of continuation events, but, in our view may distort network dynamics, for instance at change points where plenty of communities collapse in the network.

In a second example we concentrate on a single player set. A frequently occurring set in soccer matches is the 3 node set composed of two back defensive players and an attacking player of the opposite team [25]. Set 63 (players 4, 16, 19) is such a set in the match data we are using for this example. In figure 5 we show the frequency distribution with which set 63 appears and disappears, and where from and where to it continues. Each event is categorized by its type, and set formation. It can be seen that the most common events are contractions and growth from a set where player 19 is absent. Less frequently, similar events occur where player 16 is the agent of change. This distribution can inform game analysis, tactics, training and strategy. Many other type of analysis can be performed using our method, but here we are just concerned in exemplifying the method and taxonomy usage, as motivation for its application in this and other domains.

7. Conclusion

In this article we presented an approach and taxonomy to categorize community events in temporal networks. Temporal networks are pervasive in many domains and community structure analysis always generates a lot of interest, given its potential applicability. Having a standardization of concepts, terminology and analytic tools cannot but help advancing this field of study. As discussed here, the evolution of communities cannot be solely determined by changes in their topology, but must be contextualized by domain expertise. The demise of a community can be two very different events depending on the system they are representing. In this article, our taxonomy proposal is based on an adjusted Jaccard index that better reflects community lifecycle over time, especially on small networks, such as the one used in our empiric example. However, it is just one way of scoring community similarity, and can be replaced without compromising the overall method and taxonomy.

Our method works for discrete observations of the network as it evolves without the need to set a fixed frequency. Theoretically, the observation resolution could be increased up to a point where any new node activity would generate a new observation. In practice, to avoid computational overhead and information overload, it would be advisable to adapt our method to emerge only major structural events, avoiding reporting on trivial continuation events. This is left for future work.

Funding

This project was partly supported by Fundação para a Ciência e Tecnologia through project UID/Multi/04466/2019. Rui Jorge Lopes was partly supported by the Fundação para a Ciência e Tecnologia, under Grant UID/50008/2020 to Instituto de Telecomunicações.

8. Supplementary Material

Code, in the Python programming language, that implements the method proposed here-in is available at https://github.com/ramadap/Community-Lifecycle

References

[1] Sitaram Asur, Srinivasan Parthasarathy, and Duygu Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3(4):1–36, 2009.

[2] Thomas Aynaud, Eric Fleury, Jean Loup Guillaume, and Qinna Wang. Communities in evolving networks: Definitions, detection, and analysis techniques. In *Dynamics On and Of Complex Networks*, volume 2, pages 159–200. Birkhäuser, New York, NY, 2013.

[3] Thomas Aynaud and Jean-Loup Guillaume. Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks. In *Proceedings of the 5th SNA-KDD workshop*, number September 2011, 2011.

[4] Thomas Aynaud, Jean-loup Guillaume, Qinna Wang, and Eric Fleury. Communities in evolving networks : definitions , detections and analysis techniques. *Computer Networks*.

[5] Albert-László Barabási. Chapter 9: Communities. *Network Science*, 2015.

[6] Piotr Bródka, Stanisław Saganowski, and Przemysław Kazienko. GED: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14, 2013.

[7] Remy Cazabet and Giulio Rossetti. Challenges in Community Discovery on Temporal Networks. In Petter Holme and Jari Saramäki, editors, *Temporal Network Theory*, pages 181–197. Springer, Cham, 2019.

[8] Narimene Dakiche, Fatima Benbouzid-Si Tayeb, Yahya Slimani, and Karima Benatchba. Tracking community evolution in social networks: A survey. *Information Processing and Management*, 56(3):1084–1102, 2019.

[9] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 09008(9):219–228, 2005.

[10] Raju Enugala, Lakshmi Rajamani, Kadampur Ali, and Sravanthi Kurapati. Community Detection in Dynamic Social Networks : A Survey. *International Journal of Research and Applications*, 2(6):278–285, 2015.

[11] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[12] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[13] Jessica Gordon-Roth. Locke on Personal Identity. In Edward N Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 201 edition, 2019.

[14] Derek Greene. Tracking the Evolution of Communities in Dynami c Social Networks. In *2010 International Conference on Advances in Social Network Analysis and Mining*, 2010.

[15] Tanja Hartmann, Andrea Kappes, and Dorothea Wagner. Clustering evolving networks. In L. Kliemann and P. Sanders, editors, *Lecture Notes in Computer Science*, volume 9220 LNCS, pages 280–329. Springer, Cham, 2016.

[16] John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman. Tracking Evolving Communities in Large Linked Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5249–5253, 2004.

[17] Paul Jaccard. The distribution of flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[18] Manel Ben Jdidia, Céline Robardet, and Éric Fleury. Communities detection and the analysis of their dynamics in collaborative networks. *International Journal of Web Based Communities*, 5(2):195–211, 2009.

[19] Rocco Langone, Raghvendra Mall, and Johan A.K. Suykens. Clustering data over time using kernel spectral clustering with memory. *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Orlando, FL, 2014*, (December):1–8, 2015.

[20] Leroy E Loemker. *G. W. Leibniz: Philosophical Papers and Letters*, volume 53. Kluwer Academic Publishers, Dordrecht / Boston / london, 2nd edition, 1969.

[21] Raghvendra Mall, Rocco Langone, and Johan A.K. Suykens. Netgram: Visualizing communities in evolving networks. *PLoS ONE*, 10(9):1–24, 2015.

[22] Minh Van Nguyen, Michael Kirley, and Rodolfo García-Flores. Community evolution in a scientific collaboration network. *2012 IEEE Congress on Evolutionary Computation, CEC 2012*, pages 10–15, 2012.

[23] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, 2007.

[24] Luis Ramada Pereira, Rui J Lopes, and Jorge Louçã. Syntgen: a system to generate temporal networks with user-specified topology. *Journal of Complex Networks*, pages 1–26, 2019.

[25] João Ramos, Rui J. Lopes, Pedro Marques, and Duarte Araújo. Hypernetworks reveal compound variables that capture cooperative and competitive interactions in a soccer match. *Frontiers in Psychology*, 8(AUG):1–12, 2017.

[26] João Paulo Ramos, Rui J. Lopes, and Duarte Araújo. Interactions between soccer teams reveal both design and emergence: Cooperation, competition and Zipf-Mandelbrot regularity. *Chaos, Solitons and Fractals*, 137:1–7, 2020.

[27] João Ribeiro, Keith Davids, Duarte Araújo, José Guilherme, Pedro Silva, and Júlio Garganta. Exploiting Bi-Directional Self-Organizing Tendencies in Team Sports: The Role of the Game Model and Tactical Principles of Play. *Frontiers in Psychology*, 10(October):1–8, 2019.

[28] Giulio Rossetti and Rémy Cazabet. Community Discovery in Dynamic Networks: a Survey. *ACM Computing Surveys*, 51(2):1–37, 2018.

[29] Myra Spiliopoulou. Evolution in Social Networks: A survey. In *Social Network Data Analytics*, pages 149–175. Springer, Boston, MA, 2011.

[30] Yang Sun, Junhua Tang, Li Pan, and Jianhua Li. Matrix based community evolution events detection in online social networks. *Proceedings - 2015 IEEE International Conference on Smart City*, pages 465–470, 2015.

[31] Mansoureh Takaffoli, Farzad Sangi, Justin Fagnan, and Osmar R. Zaiane. A framework for analyzing dynamic social networks, 2010.

[32] Mansoureh Takaffoli, Farzad Sangi, Justin Fagnan, and Osmar R. Zaïane. Community evolution mining in dynamic social networks. *Procedia Social and Behavioral Sciences*, 00:48–57, 2011.

[33] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping community detection in networks: the state of the art and comparative study. *ACM Computing Surveys*, 45(4):1–35, 2013.