A Taxonomy of Community Lifecycle Events in Temporal Networks

Luis Ramada Pereira[1*], Rui J. Lopes[2, 3], Jorge Louçã[1],

1 ISTAR Instituto Universitário de Lisboa (ISCTE - IUL) Lisbon, Portugal
2 IT-IUL Instituto de Telecomunicações, Lisbon, Portugal
3 Instituto Universitário de Lisboa, ISCTE-IUL, Lisbon, Portugal

***ramada.pereira@iscte-iul.pt**

Abstract

Communities are one of the most important structural elements of a network. They frequently influence network behavior, which makes their identification especially useful. As a result, community detection has been a popular topic within network science in recent decades. Even more recently, fostered by an increasing availability of time stamped datasets and a pressing realization that most empiric networks are dynamic in nature, temporal networks have attracted increased attention. The time dimension introduces new network constructs and communities are not immune. A community is no longer just a bunch of fixed nodes tightly clustered, but have a life and activity of itself, shedding and gaining nodes, appearing and disappearing on the network. We believe that these dynamic constructs are still lacking a formal, consensual definition. In this article we propose a robust taxonomy of life events for communities and a rules based methodology to clearly parse these events.

Keywords: Complex systems, Networks, Clustering

1 Introduction

In static networks the ground truth of community structure is a surjection from the node set to the community set, describing community node 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 time-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 community structure to those events. This is not a new topic as it has been covered in the literature by several authors, but we believe the emerging consensus is problematic. Classifying events is not a closed problem and formalization is lacking. Furthermore, recovering lifecycle events may not be totally possible without information not inherently present in the network topology, which compounds the problem. In this article we present a simple formalized taxonomy and propose a method to track community evolution complemented by external input.

Communities are a challenging network concept. Although in this article we loosely define community 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 [3]. Extending these concepts to temporal networks obviously brings in an additional layer of complexity, which, nevertheless, has not deterred many authors from trying. In fact, expanding the ground truth of community structure to include events of a temporal nature is not a new topic as it has been covered in the literature by several authors. Barabási in his book Network Science [1] summarizes current consensus 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 not without its problems. 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 you 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, just consider the minimum quorum for a shareholder assembly or indicator species in biology.

We also find that it is easier to reason about community events anchored on the community and not on the event. So, for example, a community may experience a fragmentation while other communities in the same network may grow in size 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 into existence, a community either continues or dies. To determine continuation, we use a similarity measure, adjusted to chance, supported by an external threshold. If the similarity measure between any two communities taken from community sets at T and $T+\epsilon$ exceeds the threshold then the oldest continues in the most recent. Note that a community may continue in multiple other communities depending on their similarity. That multiplicity together with time orientation further classifies the continuation event. For example: community A_{t1} can continue in community C_{t2} and D_{t2} (a split), while community C_{t2} is a continuation of community A_{t1} and B_{t1} (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 cluster point of view, allowing for domain specific external input that further characterizes the community lifecycle.

2 Related Work

Community events have been defined by several authors [10] and 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 (communities that appear periodically). As discussed before, we think that these definitions require meta information not intrinsically present in the network topology.

One of the issues that must be addressed when determining lifecycle events is how to compare communities overtime and determine how they are related. There have been two major approaches: spectral analysis and discrete distance measures across time steps. An example of the former is [2] where the authors expand the matrix representation of a static network into a tensor by adding the time series as an additional axis and recover community structure and activity by tensor decomposition. This approach however forces the tensor size to expand to all nodes that ever existed in the network. Distance measures can save space by comparing only successive network states down to the temporal resolution of the network. Distance measures vary from ratio of shared nodes between successive timesteps [8], the Jaccard Index [7] in [10], [5], [9]:

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

and other similarity measures like in [6]

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

that favours communities similarly sized with a high ratio of common nodes. We adopted a similar approach to [5], [9], with slight modifications, while simplifying the concept of community evolution, by anchoring it on the community itself at a given point in time and not on the network. The authors in [9] propose a mechanism to automatically define thresholds without meta-information to determine community events, but it remains to be seen how closely that would follow a judgment based on problem domain expertise.

3 Recovering Community Events

Clearly defining community events is useful for many reasons, such as the development and testing of dependable temporal community detection algorithms. We need to ensure that the temporal ground truth is not open to mis-interpretation.

Our lifecycle identification approach should be able to address the problems associated with the classification of complex events when nodes exit and enter various communities as well as comprehensively cover most of the events relevant in the various disciplines where temporal networks play a role.

On this basis we created a 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 similarity exceeds an externally supplied threshold. A consequence of this rule is that remains of a community that do not reach the threshold for continuation either become a newly born community or contribute to the expansion of another.
- Depending on their multiplicity, continuation events can be further characterized:
  - From the standpoint of a community a multiple continuation event, seen from the past, is subclassified as a split.
  - From the standpoint of a community, multiple continuation events seen from the future, is subclassified as a merge.
- Expansion and contraction are sub classifications of simple continuation events with net acquisition or loss of nodes.
- Communities can die if their nodes are no longer seen on the network (death by dissolution) or because it does not continue in any other community (death by fragmentation). A community can experience loss of nodes and fragmentation simultaneously and the proper classification would then be dependent on their relative size.
- Communities can be born from new nodes (newborn) or fragments of other communities (regenerated). Both can happen simultaneously and classification follows the largest set.
- Communities can also reappear 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 "Newborn", "Regeneration", "Growth", "Contraction", "Split" and "Merge" events.

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

To compare community similarity many authors use the Jaccard Index (J) [7], as mentioned previously. [10], call it the auto-correlation function and extend it to any time delta. The Jaccard Index varies from 0, when no elements are common between communities, to $\frac{1}{3}$ when communities share half of their elements, to 1 when the communities are the same. We propose the usage of a modified Jaccard Index as described ahead.

To be able to determine life cycle events the community membership ground truth must be known at successive time steps. One of the tenets of community structure is that a random network should not have any communities (this fact is the basis of one of the most popular methods of community detection [4]).

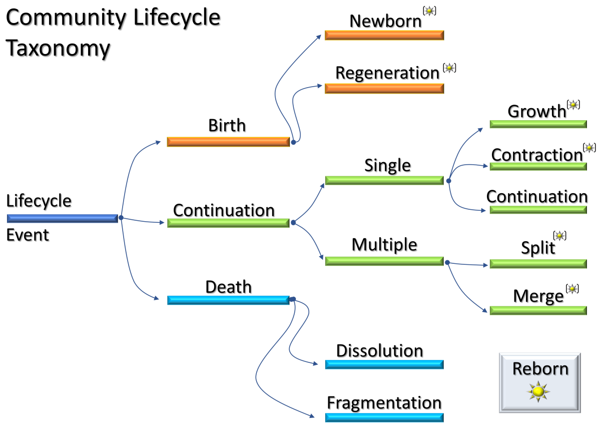


Figure 1: **Events in the lifecycle of a community in a temporal network** Classification dependent on multiplicity of continuation events and relative set sizes

However a random flow of nodes across time will result in Jaccard indexes (J) greater than zero. We note that given a multiset of community sizes at time T and $T+\epsilon$, represented by $S^{\,t}$, $S^{\,t+\epsilon}$, a random assignment of node flows results in a null Jaccard Index model $\bar{J}(C_i^t, C_j^{t+\epsilon})$ between any two communities, before discretization, given by:

$$\frac{s_i^{t+\epsilon} \times s_j^{t} \times min\left(1, \frac{\sum S^{t+\epsilon}}{\sum S^{t}}\right)}{\sum S^{\,t+\epsilon} \times (s_i^{t+\epsilon} + s_j^{t}) - s_i^{t+\epsilon} \times s_j^{t} \times min\left(1, \frac{\sum S^{\,t+\epsilon}}{\sum S^{t}}\right)} \tag{3}$$

Although discretization could have a significant impact for small networks, we believe it is still valuable to normalize to chance to extract meaning from the index (even if on very large networks the impact of index normalization is marginal), and thus we suggest the usage of an adjusted index $\tilde{J}$ as:

$$\tilde{J} = \frac{J(C_i^t, C_j^{t+\epsilon}) - \bar{J}(C_i^t, C_j^{t+\epsilon})}{1 - \bar{J}(C_i^t, C_j^{t+\epsilon})} \tag{4}$$

where $\tilde{J}$ is negative if $J \leq \bar{J}$ varying up to $1 \propto (J - \bar{J})$, with domain restricted to $1 \geq J(C_i^t, C_j^{t+\epsilon}) \geq 0$. Thus the image of $\tilde{J}$ is:

$$im(\tilde{J}) = \left[-\frac{\bar{J}(C_i^t, C_j^{t+\epsilon})}{1 - \bar{J}(C_i^t, C_j^{t+\epsilon})}, 1 \right] \tag{5}$$

Normalization has a larger impact on community pairs with higher relative size compared to the whole network. As the network grows in size and communities, random node dispersion leads to a general increase in source community diversity which lowers the null model Jaccard Index and consequently approximates $\tilde{J}$ to J, or $\tilde{J} \rightarrow J \; as \sum S \rightarrow \infty$. As an example, if we have $|c_1^t| = 200$ flowing to $|c_1^{t+\epsilon}| = 300$ in a network with 500 nodes, we have $J(|c_1^t|, |c_1^{t+\epsilon}|) = 0.5$ and $\tilde{J} = 0.34$. In a network with 5000 nodes the same flow results in $\tilde{J} = 0.49$.

The full method has the following steps:

1. A confusion (or contingency) matrix T, with size $|C^t| \times |C^{t+\epsilon}|$, is created with entries $t_{ij} = C_i^t \cap C_j^{t+\epsilon}$

2. A simple Jaccard matrix (J) is created from T and the multiset of community sizes at time T and $T + \epsilon$.

3. A null Jaccard matrix $(\bar{J})$ is created from the sequence of community sizes at T and $T + \epsilon$.

4. $\tilde{J}$ is created from T and $\bar{J}$ as described previously.

5. An external threshold θ is applied as a high-pass binary filter over $\tilde{J}$ resulting in a continuation matrix H that identifies the continuation events.

6. The row and column sum of H results in two vectors, respectively S and M that identifies a birth for $M_i = 0$, death for $S_j = 0$, split for $S_j > 1$ and merge for $M_i > 1$. The position in the matrix identifies the respective communities.

7. For every $a_{ij} = 1$ there is a continuation event between communities C_i^t and $C_j^{t+\epsilon}$ that can be simple if their size is equal and, if not, a growth or contraction event, depending on their relative size.

8. for every $S_j = 0$, we have a death by dissolution on community C_i^t if $|C_i^t| \geq 2 \times \sum_{j=1}^{|C_j^{t+\epsilon}|} t_{ij}$ or by fragmentation otherwise.

9. for every $M_i = 0$ we have a newborn event in community $C_i^{t+\epsilon}$ if $|C_j^{t+\epsilon}| \geq 2 \times \sum_{i=1}^{|C_i^t|} t_{ij}$ or a birth by regeneration otherwise.

10. The events {"newborn", "regeneration", "growth", "contraction", "split" and "merge"} can be further classified with a reborn 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 $(C_i^{t-n\epsilon}, C_j^{+\epsilon})$, where n varies from 1 to $\frac{l}{\epsilon}$ where l, ϵ stand respectively for the network longevity and temporal resolution.

To illustrate the method consider the clustering sequences $C^t = C^{t+\epsilon} = \{20, 20, 20, 20, 20\}$ at time T and $T + \epsilon$, where the flow of nodes between communities is given by the following confusion matrix:

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

This results in a simple Jaccard matrix:

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

As all communities have the same size, all elements of the corresponding null Jaccard matrix $\bar{J}$ are the same ($\frac{1}{9}$), and the adjusted Jaccard matrix becomes:

$$\tilde{J} = \begin{bmatrix} 0 & 0 & 0.25 & 0 & 0.036 \\ -0.066 & 0 & 0 & -0.066 & -0.066 \\ 0.036 & 0 & 0 & 0.036 & 0.036 \\ 0.25 & 0 & 0.25 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Let's take $\theta = 0.2$ and we get the continuation matrix:

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

Resulting in vectors $S = \{1, 0, 0, 2, 1\}$ and $M = \{1, 1, 2, 0, 0\}$. Applying the method above 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 $|C_2^t| = 20$ and $2 \times \sum_{j=1}^{5} t_{2j} = 12$, C_2^t death is by dissolution. As $|C_3^t| = 20$ and $2 \times \sum_{j=1}^{5} t_{3j} = 30$, community C_3^t dies by fragmentation. Similarly, applying point 9) of the above method, we can further classify $C_4^{t+\epsilon}$ as newborn and $C_5^{t+\epsilon}$ as regeneration.

We believe the meaning of J in the context of community lifecycle requires external subject domain information, although authors in [9] 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 matrix, it is the minimum of the maximum of the J rows and columns entries. This guarantees an increase of continuation events, but, in our view may distort network dynamics, for instance at change points where a lot of communities collapse in the network.

Another example from a synthetic network generator can be seen in figure 2. The images show only part of the whole network to highlight a mixed split / merge event. The required matrices for lifecycle determination and the resulting temporal ground truth, as output by the synthetic temporal network generator that implements the method presented in this article (code available on request), is included in figure 3 and table 1.

4 Conclusion

In this article we presented an approach and simple taxonomy to characterize community events in temporal networks. Temporal networks are pervasive in many domains and community structure 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. Although our suggested approach is based on one of many ways of comparing communities, we believe the suggested principles generalize to other approaches as well.

5 Acknowledgments

This project was partially supported by FCT (Fundação para a Ciência e Tecnologia) through project UID/Multi/04466/2019

References

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

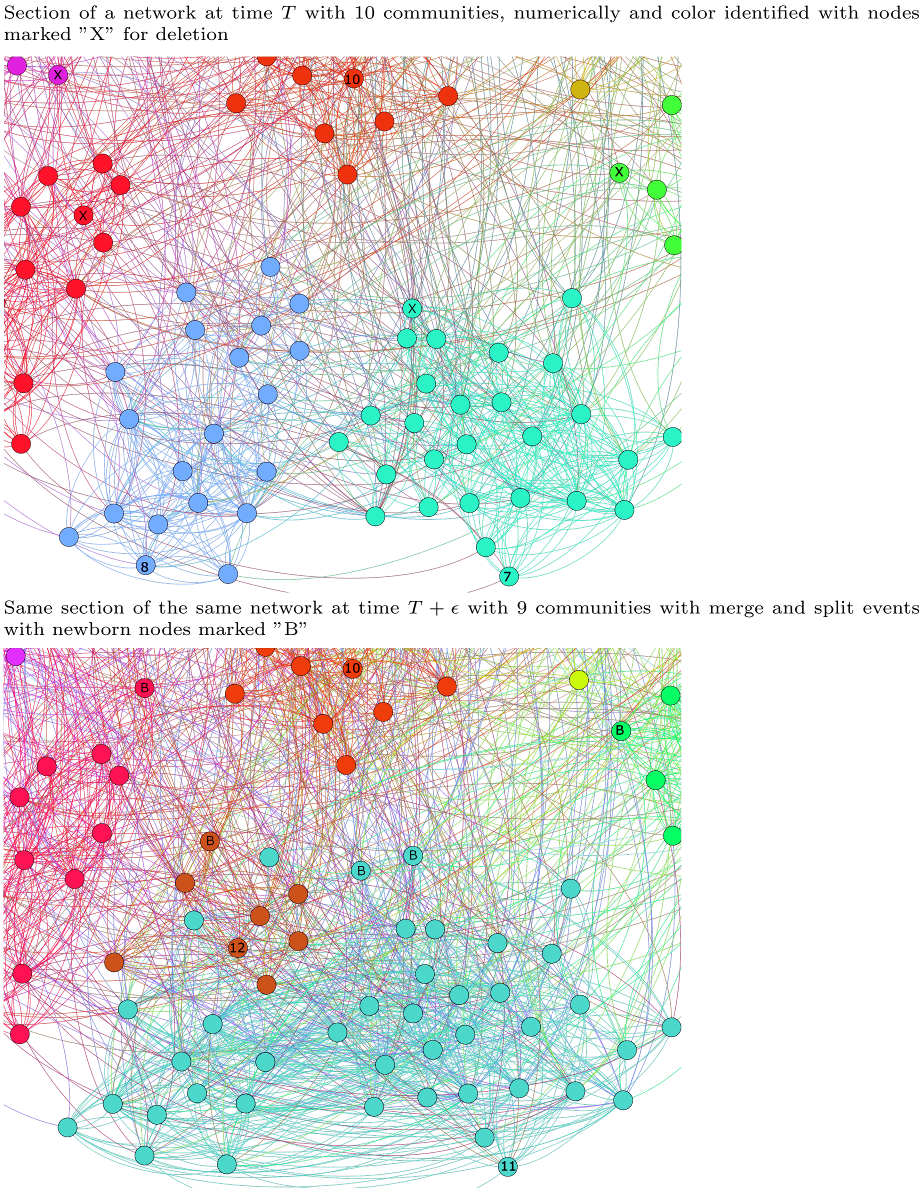
[2] Ciro Cattuto and Laetitia Gauvin. Detecting the Community Structure and Activity Patterns of Temporal Networks : A Non-Negative Tensor Factorization Approach. 9(1), 2014.

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

[4] 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.

[5] 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.

[6] 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.



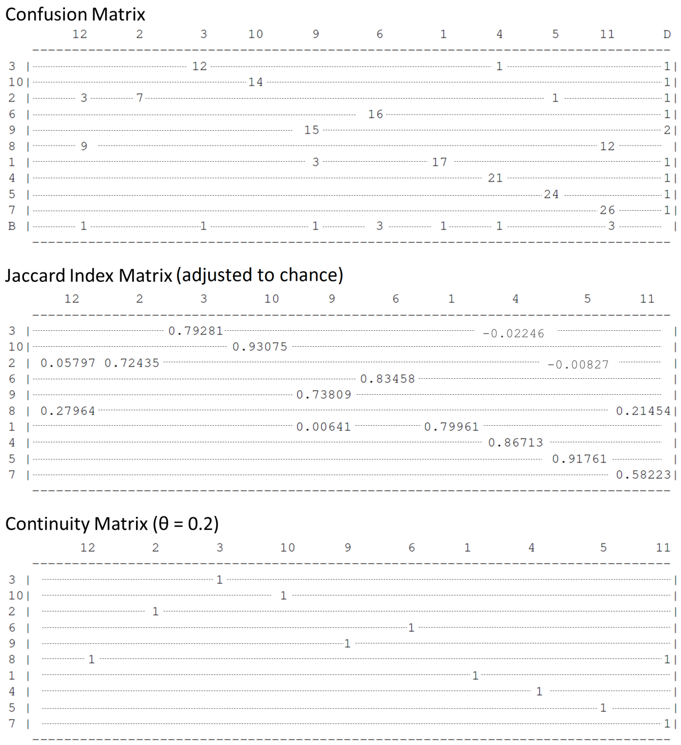
Section of a network at time T with 10 communities, numerically and color identified with nodes marked "X" for deletion

Same section of the same network at time $T + \epsilon$ with 9 communities with merge and split events with newborn nodes marked "B"

Figure 2: Community 8 split from T to $T+\epsilon$ into community 12 and split-merged with community 7 to community 11.

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

[8] Rocco Langone, Raghvendra Mall, and Johan A.K. Suykens. Clustering



Confusion Matrix

	12	2	3	10	9	6	1	4	5	11	D
3			12					1			1
10				14							1
2	3	7							1		1
6						16					1
9					15						2
8	9									12	
1					3		17				1
4								21			1
5									24		1
7										26	1
B	1		1		1	3	1	1		3	

Jaccard Index Matrix (adjusted to chance)

	12	2	3	10	9	6	1	4	5	11
3			0.79281					−0.02246		
10				0.93075						
2	0.05797	0.72435							−0.00827	
6						0.83458				
9					0.73809					
8	0.27964									0.21454
1					0.00641		0.79961			
4								0.86713		
5									0.91761	
7										0.58223

Continuity Matrix (θ = 0.2)

	12	2	3	10	9	6	1	4	5	11
3			1							
10				1						
2		1								
6						1				
9					1					
8	1									1
1							1			
4								1		
5									1	
7										1

Figure 3: Confusion, Jaccard and Continuation matrix for the sample network

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.

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

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

Community	Lifecycle Event at the end of time T
3	Continues contracting in 3
10	Continues contracting in 10
2	Continues contracting in 2
6	Continues growing in 6
9	Continues growing in 9
8	Split into [12, 11]
8	Merged Into 11
1	Continues contracting in 1
4	Continues growing in 4
5	Continues in 5
7	Merged Into 11
Community	Lifecycle Event at the beginning of time $T + \epsilon$
12	Continued from Split 8
2	Continued contracting from 2
3	Continued contracting from 2
10	Continued contracting from 10
9	Continued growing from 9
6	Continued growing from 8
1	Continued contracting from 1
4	Continued growing from 4
5	Continued from 5
11	Continued from Split 8
11	Merged from [8, 7]

Table 1: **Community events on timestep transition**
Note community 8 as it splits into 12 and merges into 11 continuing in both communities