INSTITUTO
UNIVERSITÁRIO
DE LISBOA

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

Departamento de Informática

Communities in Temporal Networks: From theoretical underpinnings to real-life applications

José Luís Loureiro Ramada Pereira

PhD in Complexity Sciences

Supervisor:

Dr. Rui Jorge Henriques Calado Lopes, Associate Professor, ISCTE-Instituto Universitário de Lisboa, Portugal

September, 2021

TECNOLOGIAS
E ARQUITETURA

FACULDADE · DE · CIÊNCIAS UNIVERSIDADE · DE · LISBOA

Departamento de Informática

Communities in Temporal Networks: From theoretical underpinnings to real-life applications

José Luís Loureiro Ramada Pereira

PhD in Complexity Sciences

Jury:

Dr. Jorge Manuel Anacleto Louçã, Full Professor, ISCTE-Instituto Universitário de Lisboa, Portugal (President)

Dr. Gergely Palla, Full Professor, Eötvös Loránd University, Budapest, Hungary

Dr. Mikko Kivelä, Assistant Professor, Aalto University, Aalto, Finland

Dr. Pedro Passos, Assistant Professor with Aggregation, Faculdade de Motricidade Humana, Lisboa, Portugal

Dr. Rui Jorge Henriques Calado Lopes, Associate Professor, ISCTE-Instituto Universitário de Lisboa, Portugal

September, 2021

Acknowledgments

For most of life's endeavors, stars must align just to consider taking them on. And a myriad of triggers need to be pulled, to bring them about. Here, I try to acknowledge examples, conditions and words of encouragement (missing many, most certainly) that lead me to this moment.

According to a survey ran by the Royal Society in the UK (The Royal Society, 2010), over 70% of people that undertake a PhD end up in careers outside of scientific research. In comparison, my path was head over heels. I spent several decades working in the Information Technology Industry, almost exclusively for the International Business Machines Corporation, commonly known as IBM. I doubt I would have taken on this project, had I worked for a different company. IBM has a culture of profound respect for pure scientific research. Collaborations with academia are paramount (IBM Research opened its first laboratory at Columbia University, over 70 years ago) and the list of notable researchers that found a home at IBM is extensive, including, among many others, scientists such as Edgar Codd of SQL fame, John Backus, the inventor of the "Backus-Nauer Form" and of the Fortran language, Charles Bennett, father of Quantum Information Theory, or Benoit Mandelbrot, the "fractalist" as he used to refer to himself. Six Nobel laureates were scientists at IBM Research. Even Albert-Lázló Barabási, of renown for his work in network theory, had a one year stint at IBM research as a post-doc. These achievements and culture had a profound influence on my interests, accrued by my first hand experience consulting with colleagues at the Thomas J. Watson center in NY, the research division headquarters, when faced with particularly difficult problems while working with some of my customers. All of this to say, that, together with the common practice of early retirement that most "IBMers" in my position usually take, freeing us at a time in our lives when it is still feasible to pursue a large gamut of interests, this professional environment was a key contributor to pursuing the project that I'm now accounting in this thesis.

So, when in 2006, freshly returned from working abroad, I started thinking about my post-business life, I got in touch with Prof. Jorge Louçã, expressing interest in joining what was at the time the Masters Program in Complexity Science. Without his acceptance, for which, and much more, I am highly grateful, I would also not be writing this thesis. There, I got my first taste of topics such as self-organization, chaos, non-linear and complex adaptative systems or agent based modeling. However, and even with full backing support from IBM, it was difficult to juggle a professional life with my academic interests, and after a rewarding and successful couple of years, these were put on hold until I could, unimpeded, dedicate the required time.

The seed, however, had been planted.

Ten years later, when I applied to join the Doctoral Program in Complexity Sciences at ISCTE-IUL, prof. Jorge Louçã was again patient enough to not only accept me, but become my doctoral supervisor. I found that his enthusiasm for complex systems studies was unabated and as contagious as a decade before. It has been, since then, a journey of discovery. Harder may be than if I had pursued it in "normal time", right after undergraduate studies. It was tough going for a while. Computer science, both theoretic and applied, was never too far away from my industry career, but more advanced topics in algebra, combinatorics, number theory, topology, probability, analysis, thermodynamics and statistical mechanics had to be re-remembered or acquired for the first time. For this I'm in debt to the professors, both from Iscte and from the Faculty of Science of the University of Lisbon, too many to name individually, whom, during the coursework of my PhD, contributed immensely to overcoming my limitations.

A special word to my colleagues Tiago, João and Álvaro, with whom I had wonderful discussions about, basically, everything, from politics to science, from Marx to Adam Smith and back, from the *pro et contra* of the research tools we employed to the challenges and solutions we faced in our academic work. Mostly over "Cozido à portuguesa", "Arroz de marisco", or other national delicacies. The COVID-19 Coronavirus disease pandemic threw a spanner in the works of many aspects of our PhD program, and these encounters have suffered as well. But guys, we should keep it going, I always found I learned something beyond how good the local cuisine is.

I probably would not have focused on network theory, and certainly not on sports science, if not for my thesis co-adviser prof. Rui Lopes. His availability, his patience to listen to my discoveries and respond with — always constructive — insights, corrections and suggestions, were unmatched. I still keep as my laptop screensaver, a photo of the extra large whiteboard in the conference room next to my office in the ISCTE-IUL basement, full of obscure expressions and diagrams after several hours of discussion about metrics, network similarity and the heuristics of optimizing a hard-problem. Prof. Rui enthusiasm for network theory and applications to real life problems were undoubtedly a driving factor for much of the contents of this thesis.

I would also like to thank Prof. Duarte Araújo from the Faculty of Human Kinetics of the University of Lisbon, for his time and constructive criticism to my work on applying network and information theory to the game of soccer. A key ingredient to having successful multi-disciplinary collaborations is having subject matter in-depth expertise and an open mind. Prof. Duarte Araújo has both in spades.

A word of thanks goes to Dr. Gergelly Palla of Eötvös Loránd University and Dr. José F. Mendes of University of Aveiro, for receiving me at their universities, and assembling a team of savvy network theory researchers to listen to my ramblings, providing perceptive commentary, and sharing some of their research with me.

I cannot finish this section without thanking my Data Science students at ISCTE-IUL, for enduring my teachings on min-cut, shortest paths, Laplacian matrices and spectral clustering,

and congratulate them for splitting the London Tube into two disconnected but balanced clusters, cutting a minimum of lines. Hopefully some of them will in the future join the community of network researchers.

Finally, I'm totally indebted to my family. To my parents without whom nothing of this kind would obviously ever be possible. But especially, to my daughter, Maria João. The "apple doesn't fall far from the tree" they say, except that in this case, you're the tree and I'm the apple. A most curious case of reverse epigenetics! I vividly recall when you gave me a private "guided tour" of the Pupin Physics Laboratory at Columbia, where Fermi first split an atom, and feeling like the proverbial doting dad, without shame, except for the tinge of envy. Your example took me here. And, by the way, thanks for proofreading my papers and for the keen use of the yellow highlighter to mark the text in "Porglish". Nothing like being brutally candid!

Tudo que existe existe talvez porque
outra coisa existe.
Nada é, tudo coexiste: talvez assim seja
certo.

Bernardo Soares (Fernando Pessoa)
Livro do Desassossego (1982)

[...] je tiens impossible de connaître les
parties sans connaître le tout,
non plus que de connaître le tout sans
connaître particulièrement les parties.

Blaise Pascal
Pensées (1670)

Abstract

Static aggregations of network activity can unravel attributes of the complex systems they represent. However, they fall short when the structure of the systems changes over time. In some cases, changes are sluggish, such as in power grids, where lines enjoy a lengthy temporal permanence. In others, a high frequency of change is observed, such as on a network of online messages, social contacts, pathogen transmission or ball passing in a soccer game. In these cases, reducing what is inherently a temporal network to a static one, leads necessarily to a loss of information, such as causal relationships, precedence or reachability rules. Temporal networks are thus the main subject of this thesis, centered on the study of network evolution from the point of view of its clusters as significant meso-structures. The thesis has two interrelated parts. In the first, theoretical challenges are addressed and original algorithms, methods and tools are developed that can further the study of network theory. In the second, these developments are applied to the analysis of team invasion sports. A measurement of game dynamics was created based on a temporal network representation of a match, with nodes clustered by spatial proximity. These measurements were found to correlate with match events of known dynamics. Moreover, they reveal unique, multi-level, aspects of the game, from the individual players contributions, to the clusters of interacting players, to their teams and their matches, which is useful for game analysis, training and strategy development.

Keywords: Temporal networks, Complex systems, Clustering, Team Invasion Sports, Dynamics

Resumo

As agregações estáticas das ligações de uma rede podem revelar atributos dos sistemas complexos que representam. Todavia, são insuficientes quando a estrutura dos sistemas se altera com o tempo. Em alguns casos, as transformações são lentas, tais como em redes de transmissão de eletricidade em que as linhas se mantêm inalteráveis por largos períodos de tempo. Noutras, regista-se uma taxa elevada de mudança, como por exemplo numa rede de mensagens em linha, contatos sociais, transmissão de patógenos ou passes num jogo de futebol. Nestes casos, reduzir o que é inerentemente uma rede temporal a uma rede estática, leva a uma perda de informação, tais como relações causais, regras de precedência ou de acessibilidade. Redes temporais são assim o tema desta tese, centrada nos seus agrupamentos, como meso-estruturas significantes. A tese está dividida em duas partes. Na primeira, são considerados problemas teóricos, e são desenvolvidos algoritmos, métodos e ferramentas que avançam o estudo da teoria de redes. Na segunda, estes desenvolvimentos são aplicados à análise de jogos desportivos coletivos de invasão. Foi criada uma medida de dinâmica do jogo, baseada na representação da partida através de uma rede temporal de nós agrupados por proximidade espacial. Os resultados obtidos correlacionam-se com eventos do jogo de dinâmica conhecida. Adicionalmente, esta medida revela aspetos únicos e multi-nível da dinâmica do jogo, desde a contribuição individual do jogador, até aos agrupamentos de jogadores, da equipa e das partidas, útil para a análise do jogo, de treino e de desenvolvimento estratégico.

Palavras-Chave: Redes Temporais, Sistemas complexos, Agrupamentos, Jogos Desportivos Coletivos de Invasão, Dinâmica

Contents

List of Tables

List of Figures

Acronyms

CM Configuration Model.

GEXF Graph Exchange XML Format.

I.I.D Independently and identically distributed.

Iscte Iscte — University Institute of Lisbon.

J Jaccard Index.

LFR Lancichinetti, Fortunato & Radicchi Benchmark model.

MI Mutual Information.

MRT Mass Rapid Transit.

NMI Normalized Mutual Information.

NVI Normalized Variation of Information.

VI Variation of Information.

XML Extended Markup Language.

Chapter 1

Introduction

Even the most impassioned reductionist would agree that some systems are way too intricate and extensive to yield to a sum-of-the-parts explanation. We are surrounded by systems, as diverse as the human consciousness, crowd psychology, ant colonies or the financial markets, whose behavior is impervious to detailed analysis. This is where complexity science, its theories, tools and techniques find purpose and utility.

The conceptual space of complexity science is vast, supported by major intellectual traditions such as dynamical systems theory, systems science, complex system theory, cybernetics and artificial intelligence. A popular map seeking to portray the history and development of complexity sciences, including major topics, themes and notable researchers, can be found in Castellani (2018). The academic field of network science is one of its major pillars. The reason is that all complex systems have constituents that form a network of relationships, interactions and dependencies, and, by using network science tools, it is sometimes possible to uncover and understand aspects of the behavior of the underlying complex system they portray.

This thesis is an account of the study of networks in the specific context of complex systems: in particular, the focus is on the meso-structures that emerge and evolve in temporal networks. These meso-structures, usually referred to as communities, clusters, groups or modules, are sets of nodes that predominantly interact among themselves. They frequently have an over-sized impact on the system's response, as found in examples such as innovation hubs in technological development, or protein-protein interactions, fundamental for functional regulation.

The problem with communities

Communities are pervasive in everyday life, and no one should have any difficulty in coming up with several examples. A family, a flock, a neighborhood, a forest, a football team, all have community attributes. They are made up of related members. These relations can be represented by a network of interactions. These are the traditional nodes and links of network theory (or vertices and edges of graph theory). If examples are simple to come by, a formal and rigorous definition is however somewhat more illusive. Intuitively and informally, communities

can be defined as sets of nodes with a higher density of links between themselves than to nodes in other sets. For example, in a soccer game, defenders from one team are more likely to interact with forwards of the adversaries than to other roles of players. This is a simple enough definition, but that hides a sea of complexity. Consider that this definition can be relaxed to define a community as sets of nodes with a higher density of links between themselves than to nodes in any other community, with no loss of intuitiveness, while the resulting community structure can become quite different (Radicchi, Castellano, Cecconi, Loreto, & Paris, 2004). On the other hand, a community may itself contain sub-communities, members that are more interconnected than others. For instance, employees of one company are more likely to interact with colleagues, than with employees of another company. However, within a single company a similar occurrence can be found inter divisions, and then inter departments and so on. Where do we make the cut?

A common way introduced in (M. E. Newman & Girvan, 2004), is to compute the difference between the fraction of the number of links that fall within the communities (or modules, thus its name "modularity") and that same fraction if the network links were randomly rewired, while keeping the node degrees unchanged. If we apply this method to all possible partitions of the network and select the one with the highest modularity, we have *likely* found the optimal network partition. This is however computationally intractable except for tiny networks, as there are B_n possible partitions of a network where B_n represents the Bell number of n. Several heuristics have been proposed to address this challenge, such as simulated annealing, spectral methods, greedy methods and many others (Blondel, Guillaume, & Lefebvre, 2008; Duch & Arenas, 2005; Guimera & Amaral, 2005; M. Newman, 2006) with varying success. We emphasized *"likely"*, as modularity optimization has a resolution limit, merging small communities into larger ones, even if they are well defined, when the network is sufficiently large (Fortunato & Barthélemy, 2007). This represents a particular challenge to community identity in temporal networks, as a community can stop being detected, even if it has not changed, when the network grows somewhere else. Other methods, unrelated to modularity optimization have been introduced, but this is still an area of active research.

Community definition gets even harder when considering that in many empiric systems nodes can belong to multiple communities. This has led to a specific line of research dedicated to studying overlapping communities. A higher order of complexity is also introduced when considering that communities change and, depending on their connection activity, a time independent definition of community may be challenging, if not impossible. In real life, we find communities that keep their identity even when all of its members have been replaced: consider the supporters of a centuries old soccer club[1], where their original members are long gone, but the community of supporters is still very much active today.

[1] Sheffield F.C. the oldest club in existence was founded in 1857

Community detection, identity and ground truths

These issues are of particular importance to community detection, if we assume that the community structure is somehow encoded in the network. In this thesis, we are majorly concerned with community identity and evolution. If all we have is a record of nodes and links, and the network is static, i.e. "frozen in time", detection and identification of communities are basically synonymous. As our focus is temporal networks, we make a nuanced distinction between detection and identity. We can detect a community C_1 at time t, we can detect a different community C_2 at time $t_{+\delta}$ if its members have changed, but its identity may be kept. Think about a family with a new born baby. This is one reason why the resolution limit discussed above can be so damaging.

Although the "fuzzyness" associated with the concept of community as exposed above is a recognized difficulty, in this thesis we are not directly concerned with cluster or community detection. Our work starts from the partitioning of a network into modules, in what is commonly known as the ground truth of community membership.

Research questions

Encapsulating in a single sentence the multi-year effort of the research work for a doctoral program is never easy. One easily gets lost in the meandering of threads that inescapably come up as inherent to the actual research process. However, in the case of this thesis, the question "How can understanding and quantifying cluster evolution in complex networks contribute to insights into the systems they represent?" is fit for purpose.

The motivation for this research question stems, most importantly, from the recognition that work on complex networks is a core pillar of complex systems studies, and its breadth of potential applications is ever expanding. Most, if not all, real-life systems are systems that evolve in time, so the field of temporal networks will not be empty of research opportunities any time soon. Communities, as argued in the previous section, are one of the most important substructures of networks, so covering both seemed logical and synergetic. The recognition that, especially for higher order complex networks, there is still an extensive whitespace for new contributions, was a complementary motivation.

From this main research question, several sub topics emerged that drove the research activites and related publishing:

- how to create temporal datasets with known structural properties to test hypothesis? If there is a need to test hypothesis on properties such as the degree joint probability distribution, or cyclic inter node activity, or change points in a network, sometimes empiric data is not readily available, especially if space scanning is needed for one or several of these properties (see (Nicosia et al., 2013) for a comprehensive discussion of unique aspects of temporal network measures).

- how to characterize the lifecycle of a community? how can we address the problem of community identity, or in other words when is a community no longer recognized as its former self?
- how to apply and validate methods against empiric complex systems?

After a critical review of some of the current knowledge about these and related topics, I try to answer these questions in the following three chapters. These chapters were previously published as Pereira, Lopes, and Louçã (2020, 2021); Pereira, Lopes, Louçã, Araújo, and Ramos (2021)[2].

Structure of this thesis

After this introduction, in chapter 2 we complement the "related work" sections (3.2, 4.2 and 5.2) of the three published articles, with a critical review of transversal temporal networks topics published in the scientific literature.

This chapter is then followed by chapter 3, where we describe a system to generate temporal networks, using a multilayer formalism under user parametrization of topological features. Although several such methods and systems exist for static networks, their number for temporal networks is very limited and we could not find one with the flexibility and principles that we considered important. Several challenges of developing such a system were addressed and original solutions created. One of the key challenges was finding the minimum amount of change that a network can experience after re-partition. This allow us to split the measurement of the network changes into two components, first, the changes "forced" by the re-partitioning, and second, the change emerging from additional node re-labelling. We made an initial, tentative, exploration of these two measurements in the context of the soccer game as reported in chapter 6.

Reflecting the relative "infancy" of temporal networks studies, there are no real commonly adopted standards to represent a temporal network in machine readable format. For Syntgen, we opted to use GEXF, an XML based format that seems to have wide support in network analysis tools, such as Networkx (Hagberg, Schult, & Swart, 2008) or Igraph (Csardi & Nepusz, 2006).

Chapter 4 proposes a taxonomy of community lifecycle events that goes beyond what have been so far proposed in the network theory literature. The value of tracking the lifecycle of communities or clusters is unquestionable in areas such as marketing, sociology, biology, security, and so on, basically everywhere where communities exist. Examples are trivial. We apply the approach we developed to an empiric network representing a soccer game, starting the applied phase of the work described in this thesis.

In the chapter 5 we apply temporal networks to analyze a competitive field invasion team sports, soccer, represented as a multislice hypergraph. We used information theoretic constructs

[2]These chapters may include minor textual and format adaptations to the published articles, for proper integration into the style and contents of this thesis.

introduced in chapter 3 to measure the game dynamics and found strong correlation with notational metadata. This approach is sufficiently generic to extend to other complex systems with temporal attributes, where a distance metric between layered observations of cluster labeled nodes can be a revealing measure.

We conclude in chapter 6, discussing limitations of our work, possible ways of addressing them, and follow-on, future work, inspired by the research accounted in here.

Terminology

Finally, a word about terminology. Graphs and networks are related but not necessary synonymous. We define networks as a representation of a complex system using graphs. However, terminology from graphs and networks are frequently switched, and we may not be totally free from indulging in this practice, even though we are very much aware that it does not help reducing the "semantic fog" that afflicts network theory. The fact that this is usually the case in emerging sciences, should not give us pause. In the meantime, although we strive to be contextually appropriate, we may lapse to use the entries in the following tuples interchangeably: (link, edge), (vertex, node), (graph, network, and their derived compounds), (cluster, module, community), and (clustering, partition, node labeling).

Chapter 2

Temporal networks in the scholarly literature

The study of graphs and networks has come a long way since the very first paper published by Leonard Euler on the Seven Bridges of Königsberg in 1736. Major contributions were proposed in the mid 20th century in the work of Paul Erdös and Alfréd Rényi. These authors introduced the classic Erdös-Rényi model of network generation (Erdös & Rényi, 1960), where nodes are randomly attached according to a given probability p, departing from the regular lattice-like model of the traditional graph theory. Attributes of the networks built with this generative model were extensively studied, and it soon became apparent that most empiric systems did not conform to the emerging topologies that they exhibit. New contributions, such as the famous "Small world" model of Watts and Strogatz (1998), where, starting from a regular lattice, nodes are rewired with probability p, or the "preferential attachment" model (Barabási & Albert, 1999), where new nodes are wired to existing nodes with a probability that is a function of their degree, generate networks that better reflect some of the attributes seen on real life systems.

In spite of their recognized usefulness, the static networks these models generate, cannot directly capture properties of complex systems that change with time. For this, we need temporal networks[1]. An example of the limitations is the transitivity property, that does not generalize from static to temporal networks. In a static network if node A is connected to B that is connected to C, then there is a path, albeit indirect, from A to C. In a temporal network, there may exist a link from A to B and from B to C but if the link B, C precedes and never coexists with A, B, then there is no path from A to C. This is a direct consequence of the fact that constituents are not necessarily persistent in a temporal network, and that their lifespan can vary. The micro-level elements, nodes and edges, can appear, disappear and reappear, impacting the overall network and its meso-structures, such as communities. Their lifespan can go from instantaneous to permanent. For instance, in e-mail networks the link can be considered immediate. In a phone network, calls have usually a short duration. In social networks friendships

[1] Temporal networks are also referred to as evolving, dynamic or time-varying, by many authors.

are protracted. In the other extreme, in an urban mass rapid transit (MRT) infrastructure, new links and nodes emerge slowly and usually become persistent or long-lived.

Several new formalisms have been introduced in response to the need of modeling real-world phenomena and systems that cannot be adequately described by a single static network, with the branch of multilayer networks becoming increasingly popular. Multilayer networks encompass subcategories such as multislice, multiplex or networks of networks. We have found the terminology for these higher-order formalisms inconsistent and sometimes contradictory and standardized naming conventions lacking (an attempt at standardizing constructs and their terminology related to these higher order networks can be found in (Kivelä et al., 2014)).

Network dynamics and dynamics on networks

For clarity, when studying temporal networks our focus was network dynamics (or dynamics of networks, as referred to by some authors) and not necessarily in dynamics on networks (Mukherjee, Choudhury, Peruani, Ganguly, & Mitra, 2013). Using the MRT as illustration, we can have a static network of stations (nodes) and tracks (links) over which a dynamic process of people movement takes place. Percolation and spreading phenomena, such as information diffusion or epidemics are similar examples of dynamic processes that can be studied over a static network. In the subject networks of this thesis the actual structure evolves over time. Some authors consider that a temporal network "moves information about when things happen from the dynamical system to the network, the underlying structure on which the dynamics happen" (Holme & Saramäki, 2012, p.99), we however keep on seeing this as two independent processes: the fact that the network itself experiences dynamical topological processes (the MRT adds new and decommissions old stations and lines) does not preclude the existence of another dynamical process on top of the temporal network. We can illustrate this assertion with the simple example of pathogen transmission on a temporal network representation that encodes physical contacts. If A, B precedes B, C, and A infects B, the possibility of transmission to C is not guaranteed. At the network level, there is indeed a path from A to C, but at the dynamical system level running on the network (the disease spread) it depends on whether B is still infectious when its link to C emerges. In summary, contagion is dependent on the dynamical system on the network, and only subsequently on the network dynamics. Another example could be the network representation of a chemical reaction, where there must be coexistence of some reactants for the reaction to proceed.

When searching the scholarly literature for network dynamics, one comes across many writings on network evolution (Vázquez, 2003); (Dorogovtsev & Mendes, 2002); (M. Newman, 2018, p.434); (Leskovec, Kleinberg, & Faloutsos, 2007); (Barabasi, 2016, Chapter 6); (Barabási & Albert, 1999); (Jeong, Néda, & Barabási, 2003). Their focus is, in fact, on network formation and their emerging properties, covering models such as Price's preferential attachment, Barabási-Albert model, Bianconi-Barabási model, Watts-Strogatz model, and others.

Here, we are not addressing the inquiry of evolving networks as a mean of studying the processes that lead to their "final" observed state, that is, we are not trying to inquire about the process of topology emergence, but rather observe, classify, and explore changes in networks, as they occur, especially clustering changes.

Temporal network definitions and representations

Within the disparity of terminology, sometimes concepts and constructs can become fuzzy. This is certainly the case with temporal networks. We don't recognize temporal networks as a topological formalism, such as a bipartite graph, an hypergraph or a multiplex network. This difference goes beyond the commonplace definition of a network as a graph based representation of a complex system. In practice many topological formalisms can be used to represent a temporal network. In (Kivelä et al., 2014, p.206), authors identify 26 different types of networks found in the literature, many of which could be used to represent a temporal network. Tellingly none of these types is a "temporal network". We consider that the "temporal" in "temporal network" refers, first and foremost, to a property of complex systems that suffer evolution over time and not necessarily to any specific network or graph framework.

That is not to say, that temporal networks do not require appropriate formalisms for representation. They can be represented by a multilayer network with every layer corresponding to a timed observation of the network. These are usually aggregated observations over a time interval (Δ_t), and as such are a discretization of what are typically continuous phenomena. This introduces error, that can be minimized by reducing Δ_t, which increases layers sparsity and cardinality. This is represented in figure 2.1 where a stream of instantaneous links, such as instant messages, e-mails, financial payments and others, is sampled at varying time intervals. In the limit, when $\Delta_t \to 0$, we recover a continuous domain, albeit with only simultaneous edges per layer. Determining the appropriate time interval is problem domain dependent (Latapy, Viard, & Magnien, 2018, p.19). In this formalism, directed edges couple adjacent layers, conditioned by the arrow of time, linking the same node and signaling node survival. Inter-layer links only connect the same node. Some authors classify this as multiplex networks (Gómez et al., 2013). Others, make a distinction, based on the property that the inter-layer links are node-aligned and ordinal, which is not a general requirement for a multiplex network. For instance the author in (Bianconi, 2018, p.80), call this type of formalism a multi-slice network. In reality a multiplex network is no more than a multigraph with colored edges, where the color describes a given interaction property, such as the role of a social agent, or the medium of a communication, or the vehicle type in a transportation network. Each color in the space of colors is a layer and nodes can connect across layers, and have no inherent order, they can be purely categorical. In this case, they do not need to abide by the adjacency or directional property, potentially creating a clique of a given node occurrence across multiple layers, signaling identity but not lifespan. Other authors still, defend that a temporal network may require a non-multiplex multilayer rep-

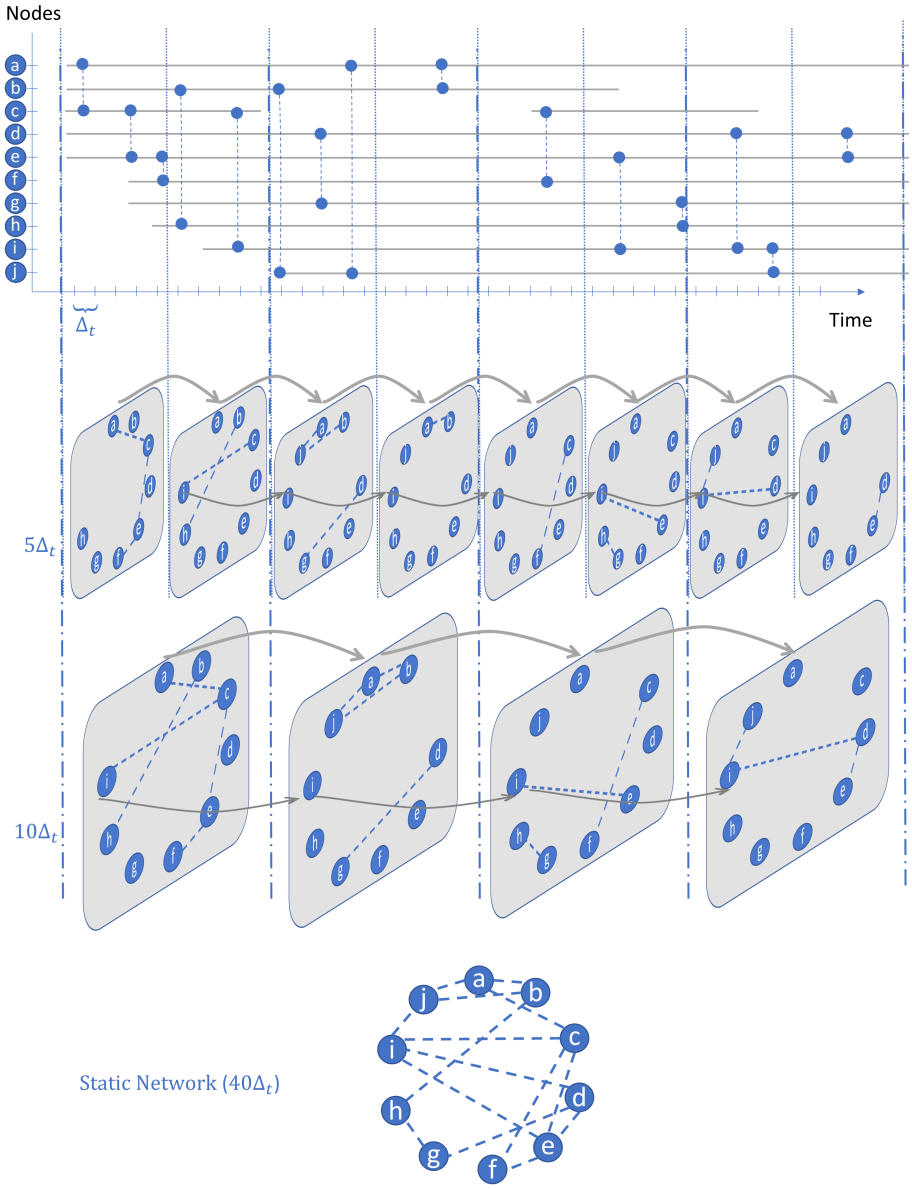


Figure 2.1. In this example, a temporal network with 10 nodes is discretized into a multi-layer, time sliced network at different resolutions ($5\Delta_t$ and $10\Delta_t$). The inter layer arcs, represented in gray (for readability, only included for nodes "a" and "i"), denote that the layers are node aligned. Extending the sampling intervals increases the loss of temporal information. This loss hits a maximum when collapsing the link set into a static network, erasing all temporal information.

resentation (De Domenico et al., 2013, p.4), that is, connections may exist between different nodes in different layers, for instance to represent causal, time-mediated relationships.

Multi-layer networks can be extended by so-called aspects (Kivelä et al., 2014; Pilosof, Porter, Pascual, & Kéfi, 2017, p.208), further contextualizing the space of relations between nodes, that is, one aspect could be the time of an interaction, another aspect the medium, and so on. Aspects differ from layers in that they refer to the same event/interaction. They can be viewed as different attributes of a single link.

Many of these representations extended the graph adjacency matrix to a tensor, where the

toolkit of tensor decomposition has been used to reveal many properties of the represented networks (Kolda & Bader, 2009; Martin & Porter, 2012), albeit at a cost of forcing two of the tensor dimensions to have the same size as the number of nodes that will ever exist in the network, i.e. padding the tensor with null degree nodes. One, however has to be careful, as in this case a null degree node is not the same as an isolated node, and some metrics may need to be adjusted. Additional aspects are represented by additional tensor dimensions, although it is possible to reduce dimensions, flattening the tensor, potentially all the way down to a 2D matrix, called a supra-adjacency matrix (Cozzo, Ferraz de Arruda, Rodrigues, & Moreno, 2018; De Domenico et al., 2013; Kivelä et al., 2014; Solé-Ribalta et al., 2013), by expanding nodes accounting for their presence in aspects and layers.

Real-time streaming data and change points

Temporal networks built out and processed in real time from streaming data, suggest a different type of formalism, as the total number nodes is usually unknown. A time series of contacts or adjacency lists may be a more suitable representation.

While previously mentioned representations of a temporal network are information preserving down to their time resolution, it is possible, to create other types that sacrifice some of the (hopefully less relevant) information to reduce the space complexity of perfect memory. The author in (Holme, 2015, p.7-9) introduces several different approaches, such as using link-weighted graphs, with weights counting edge occurrences, Markov process transition matrices, or concurrency graphs.

A typical problem facing complex systems researchers is the detection of change points in temporal networks. For streaming (or longitudinal) data, authors in (McCulloh & Carley, 2011) propose a method for change detection based on statistical process control using control chart schemes that signal change. They found the cumulative sum control chart (Page, 1954) to be effective against one simulated and three empiric networks.

Authors in (Peel & Clauset, 2015) generalize the hierarchical random graph model, basically a non-binary tree dendogram, specifying edge probability at multiple levels, and use it to define a distribution over networks. After fitting edge probabilities for the network in question, change points are detected when a threshold conditioning false positive rate is exceed over a given time window. They apply this technique to synthetic data and to two empiric networks, the MIT Reality Mining proximity network and the Enron e-mail network, and find good correlation with known external events.

Authors in (Darst et al., 2016) propose a method to compute timescales, that is periods of time when a system is relatively stable, by using the Jaccard Index, a similarity measure, to compare successive periods, increasing the number of observations until the index tendency turns negative. The intuition is that when computing the ratio between the intersection and the union of the set of events of a previously identified period and a set of successive later events,

it is likely that initially the index will grow as the probability of finding events in this set that have already been observed is high, but it will get to a point where new unobserved events will outweigh those, inverting the index tendency. At this point a new period can begin. This could be useful for aggregating observations of a network when a time sliced representation is sought. Similarly to the previous reference, the authors test their approach against empirical networks (Enron e-mail database, MIT Reality Mining, twitter dataset of tweets with a common hashtag), for which external events are known, and find strong correlations.

Applications of temporal network theory

Temporal networks have found applications in multiple domains. In this thesis, we studied team invasion sports, and a comprehensive discussion of related prior work is included in section 5.2 of the article that makes up the chapter 5. Here we mention some other instances where temporal networks have been of use.

In many articles, original contributions to temporal network theory are exemplified through their application to network representations of complex systems. Some of these systems have publicly available datasets, which contributes to their popularity. Cases in point are the Enron accounting scandal e-mail database (Cohen, 2015) and the MIT reality mining proximity network, a social network of mobile phone bluetooth traced contacts (Eagle & Pentland, 2006). Examples of articles using these datasets to validate and exemplify proposed methods of analysis were referenced in the previous section (Darst et al., 2016; Peel & Clauset, 2015). Citation networks extracted from repositories such as the Web of Science are also popular. Citation networks have been shown to exhibit preferential attachment properties (M. E. Newman, 2009). In (Medo, Cimini, & Gualdi, 2011) authors use citation data from the American Physical Society to propose a network generation model that combines heterogeneity and temporal decay into the node fitness of the preferential attachment model, two properties that have previously been shown in citation networks, but in isolation (Bianconi & Barabási, 2001).

Other articles are more of an applied nature, using publicly[2] or privately available datasets. An example can be found in (Génois & Barrat, 2018), where authors question the usage of co-location records, automatically and wirelessly captured, as a proxy for face-to-face contacts. They used several datasets, collected in facilities of health providers, in schools and in conference settings, and find that some features of the underlying system can indeed be recovered, when comparing a down-sampling of the co-presence records to the ground truth of physical social contacts, independently observed. They then apply it to the identification of containment strategies of epidemic processes, deriving guidelines for the usage of automated collection of

[2]Other public network repositories that store empiric temporal networks are available from Stanford University (https://snap.stanford.edu/data/), that includes mostly data from digital social networks, and from a collaboration of the ISI Foundation – Turin, Italy, the CNRS – Centre de Physique Théorique – Marseille, France, and Bitmanufactory, Cambridge - UK (http://http://www.sociopatterns.org/) where multiple datasets can be found containing real life networks of co-location of humans and of other animals captured by a sensing platform.

co-location events in these contexts. Another example, but involving the tracking of enclosed Baboons (Gelardi, Godard, Paleressompoulle, Claidiere, & Barrat, 2020), finds similar results, which opens the possibility of deployment of the sensing platform and of the temporal network methods of analysis to wild animal populations, otherwise difficult, costly and cumbersome.

Using citations and co-authorships from the arXiv repository, authors in (Amblard, Casteigts, Flocchini, Quattrociocchi, & Santoro, 2011) show the advantage of using temporal networks in detecting structural changes and evolution of the underlying networks, and how co-authorship and citations co-evolve.

Other publications on temporal networks

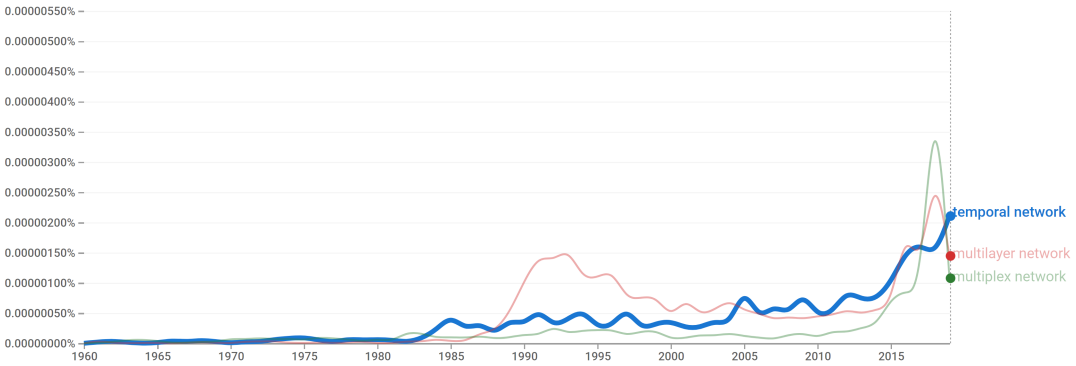


Figure 2.2. Higher order networks have become a popular topic in recent decades. Source: *books.google.com/ngrams*, for years 1960-2019

Although temporal networks are a very trendy topic (see figure 2.2), it is still a small proportion of the scholarly corpora of network theory. Books on general network theory rarely devote more than a few lines or pages to the topic of temporal networks (Barabasi, 2016; Biagini et al., 2019; Latora, Nicosia, & Russo, 2017; Lewis, 2009; Menczer, Fortunato, & Davis, 2020; M. Newman, 2018). Dorogovtsev and Mendes (2002) covers evolving networks but from a generative perspective, i.e. growth and structure. Most of the scientific literature on temporal and multilayer networks is found in journals and in edited books. Of the non-edited books, the reader is referred to Bianconi (2018), a textbook fully dedicated to multilayer networks covering formalisms, measures and dynamics, and to Cozzo et al. (2018) a book dedicated to multiplex networks, discussing the mathematical constructs that are foundational for high order complex network analysis. Specific to temporal networks, P. Holme and J. Saramäki edited two books (Holme & Saramäki, 2013, 2019) providing multiple authors perspectives about the theory and applications of temporal networks, many of which have been cited in this thesis. These two authors also published extensive articles covering the state of the art on temporal networks (Holme, 2015; Holme & Saramäki, 2012).

Research gaps and concluding remarks

As mentioned in chapter 1, the literature review included in this chapter is complementary to those included in chapter 3, 4 and 5 and does not explore in detail the specific topics that those chapters address. In particular, in section 3.2 we review the literature on synthetic network generators, their applications and the theoretical aspects that they invite. In section 4.2 we discuss the contemporary scientific understanding of community lifecycle and previous proposals for their determination, and finally in section 5.2, we review the literature focused on the application of network theory to sports science, especially to team invasion sports.

Here, we limited our inquiry to critically investigate and pinpoint challenges related to temporal network theory constructs. We found gaps on collective acceptance of what some of these constructs are, and this has guided us when introducing formalisms in subsequent chapters. It is clear to us that the field of representing time evolving complex systems as temporal networks, be they a social enterprise, a biological ecosystem, the emergence of phenotypes in an organism, or the 90 min of a soccer match, is far from exhausted, and is open to multiple threads of research. We hope to have contributed to a journey that, in the time scale of scientific pursuit, is just beginning.

Chapter 3

Syntgen: a system to generate temporal networks with user-specified topology

The following section was previously published in (Pereira et al., 2020). A researcher wanting to study network communities and their lifecycle faces the challenge of obtaining appropriate samples that conform to specifications appropriate for his research questions. As an example, if he wants to study how joint degree distribution influences the emergence of community structure, he needs samples with controlled joint degree distributions. The same can be said for other attributes, such as density, size, path length, clustering coefficient, vertex eccentricity, modularity and so on. As discussed previously, community detection methods with absolute accuracy do not in general exist, thus, he is also faced with the problem of establishing a ground truth of community membership. The availability and ease of acquisition of time stamped representations of empiric networks, with known ground truth, have recently improved, but our researcher is still faced with significant challenges to control for meaningful variables.

The need to address these challenges, was the reason why, creating a synthetic generator under user-specified parameters, was the first step in the doctoral project that this thesis documents. A few approaches have been previously proposed to generate synthetic temporal networks that conform to static topological specifications while in general adopting an ad hoc approach to temporal evolution. We believed there was still a need for a principled synthetic network generator that conforms to problem domain topological specifications from a static as well as temporal perspective. In Syntgen we built such a system. The unique attributes of this system include accepting arbitrary node degree and cluster size distributions and temporal evolution under user control, while supporting tunable joint distribution and temporal correlation of node degrees. Several theoretical contributions were developed, including the analysis of conditions for graphic sequences of inter- and intra-cluster node degrees and cluster sizes, and the development of a heuristic to search for the cluster membership of nodes that minimizes change when evolving the network under constrained stochastic conditions. The system and its contributions were later used to classify community events, and study soccer, as an empiric system that can be represented by a temporal network of evolving clusters.

3.1 Introduction

Networks are all around us: computer, telecommunication, biological and social systems are just a few examples of systems of entities that interact and relate to one another in some specifiable way, producing identifiable phenomena. Graph theory, which had its origins in the 18th century when Leonard Euler published his "Seven Bridges of Königsberg" problem and its negative solution (Euler, 1736), is the basis of the field of study that has become network science. Network science is concerned with understanding networked systems, describing their micro, meso and macro scale attributes and helping us predict their behavior. Many networks exhibit groups of nodes that are more closely interconnected amongst themselves than with the rest of the network. These groups, referred to as clusters in graph theory or communities in network science, are usually of particular interest to network researchers. They may have an over-sized impact on the network behavior and their identification is often highly useful.

From its origin in graph theory, network science has focused on static networks, that is, networks "frozen in time" with link permanence. However, real world systems are rarely static: links on webpages are added and removed everyday in the world wide web, amino acid interaction for protein folding occurs over time, friendships are created, age, wither and renew. This realization led to major efforts to extend existing science into temporal networks, with several authors proposing approaches that embed time specific attributes. Communities are no exception, and several constructs have been proposed to characterize the way a community develops over time. Although some of these constructs are problematic since they cannot be derived solely from the network structure, they serve as a base that allow us to build a commonly accepted vocabulary that helps advance this field of study.

Time stamped data of empiric systems with known ground truth about communities does not abound. As extensively discussed elsewhere (M. E. Newman, 2004) even the concept of community membership is not without its challenges. This makes it more difficult to test systems that effectively recover community node membership over time. Having a system that generates a temporal network under user specified topology, with known ground truth, can help alleviate these challenges. Syntgen[1], as described in this chapter, is such a system.

Syntgen intent is to generate temporal networks that exhibit attributes observed in empirical networks. These attributes include the temporal degree correlation and the joint degree distribution. Temporal degree correlation can be seen in social networks where some nodes have continuous high popularity or, in communication networks, where nodes having periodic activity are common. The joint degree distribution expresses the probability of node links as a function of their degrees, leading to varying levels of assortative mixing as seen in many real life networks.

The input required by Syntgen at each time step is a multiset of community sizes and a bijection of two n-tuples representing sequences of total and intra-community node degrees.

[1] code available at https://github.com/ramadap/Syntgen

Optionally the user can specify preferences for joint node degree distributions, temporal node degree correlation and a set of nodes to be eliminated at each time transition.

We developed a method which Syntgen uses to test if user specifications are graphic (Behzad & Chartrand, 1967) and, if successful, generate a compliant temporal network. The user does not specify node membership or links, these are generated by the system. As there is randomness in the process of network construction, both on node community membership as well as in the network wiring, the same specifications will not typically generate the same network. However, they should asymptotically converge to the same average topology.

The user can loosely control the dynamics of the network by changing its input at chosen time steps, with new nodes created and others killed to satisfy input specifications. Changing correlation and joint distribution parameters will also impact the wiring of the network.

We provide example sequence generators that sample power laws, exponential and binomial distributions, all of which have been found in empirical networks (M. E. J. Newman, Strogatz, & Watts, 2000). These generators include parameters that specify community maximum and minimum size, maximum and minimum node degree, distribution rate parameters and a ratio (r) of intra to total degree, which can be fixed or Bernoulli distributed with $\mathbb{P} = r$. The user can use or adapt these generators or provide their own. Obviously, although the system will assign nodes to communities, these are only meaningful if the ratio of intra links to total links is sufficiently high. This ratio varies depending on the network structure and on the cardinalities of the communities. Larger communities are less stringent with their requirements. A thorough discussion can be found in (Fortunato & Hric, 2016, p.11).

With this input, Syntgen outputs a temporal network with known ground truth of its community structure for every time interval. To minimize network changes beyond those specified by the user, Syntgen tries to determine the node community membership across time steps that results in the shortest shared information distance between clusterings. An exact solution is intractable, and we develop an appropriate heuristic.

Our system works for simple networks. Syntgen generates temporal networks with no self loops or multi-links, non weighted and undirected, with no community membership overlap, with no isolated nodes, in snapshot mode. A new instance is generated at each time step and the overall temporal network is the sequence of generated snapshots. It would be possible to extend this model to a truly continuous streaming network, although a principled approach for node and edge activation would need to be devised to enforce node degree and community size affinities.

We believe Syntgen, as a temporal synthetic network generator, is unique in creating networks with arbitrary community sizes and node degree distributions and providing ways to control node joint degree distribution and node degree temporal correlation.

In the remainder of this document we review in section 3.2 other work related to the function and objective of our system. In section 3.3 we start by describing the general flow of the system, its modules and functionality, followed in section 3.3.2 by a detailed description of the approach

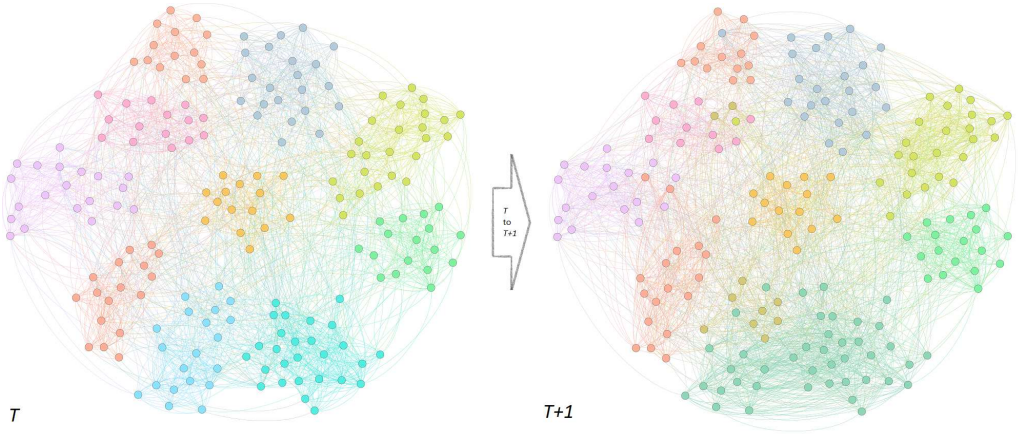


Figure 3.1. Time consecutive slices of a dynamic network as generated by Syntgen
This example is color coded according to community and has ≈ 200 nodes, $10 - 9$ communities, with multiple community events. A full description can be found in section 3.3.2.5.

we took to generate a network snapshot respecting the user topology. How we aim to reduce spurious noise when evolving the communities at every timestep is covered in section 3.3.3, and we conclude with 3 additional sections covering experiments, conclusion and future work.

A note on terminology conventions: In this subject area a vast array of terms are used to describe very similar concepts, like communities vs clusters, or partitions vs clusterings. Throughout this document, we adopt the following terminology and symbol conventions:

- ”Community” refers to groups of nodes more tightly connected amongst themselves than to the rest of the network (in lieu of terms like cluster, or partition)
- ”Clustering” refers to the splitting of a network into communities (in lieu of partition). See formal definition in 3.1.
- ”Temporal” is an attribute of a network that changes over time (in lieu of dynamic or evolving)
- ”Nodes and Links”: Links are connections between nodes at the same time step. Nodes only exist if they connect. There are no isolated nodes in Syntgen.
- We call the movement of nodes between communities across time steps, ”Node flow”.
- We define ”graphic” as the property of sequences of community sizes and bijective node total and intra degrees that enable their representation as a graph.
- We denote sets by an uppercase letter and individual elements by the corresponding lower case letter, optionally subscripted for identification. Frequently used sets and variables have their own dedicated symbol as per table 3.1.

Table 3.1. Symbol convention

Symbol	Definition
C	Clustering or community set, optionally with a subscript to indicate time step
D	Degree sequence (or ordered total degree sequence, depending on context, bijection with E, optionally with a subscript to indicate community membership).
E	Intra degree sequence (bijection with F), optionally with a subscript to indicate community membership.
F	Inter degree sequence , with $n = \|D\| = \|E\| = \|F\|$ and $f_i + e_i = d_i \ \forall(1 \leq i \leq n)$, optionally with a subscript to indicate community membership.
G	Network, optionally with a subscript to indicate time step
L	Link set
O	Kill node set (user specification)
S	Multiset of community sizes
T	Time step
U	Flow of nodes between time steps
V	Node set

3.2 Related Work

Work related to Syntgen falls into two categories:

- network science, theorems and algorithms that supported the development of our system
- prior systems that have been developed with similar or related desiderata.

In the first category we cover clustering similarity and community lifecycle events, and in the second, benchmarks for community detection algorithms and other temporal community network generators.

3.2.1 Static benchmarks for community detection algorithms

Authors in (Lancichinetti, Fortunato, & filippo Radicchi, 2008) have drawn our attention to the fact that community detection algorithms that perform well in a given network topology may be less accurate in a different topology. Prior to their work, algorithms for community detection were validated mostly against the Girvan-Newman benchmark (M. E. Newman & Girvan, 2004), which is based on a stochastic block model that only deviates from a typical random Erdős–Rényi model by the introduction of a tunable parameter specifying the probabilities of intra and inter community links, transforming the network from a pure random network to a random network of random networks (the communities). From experience we know that empirical networks do not generally follow this model. Quite often they exhibit long tail distributions of node degrees and community sizes, as is the case of networks where new nodes link with higher probability with existing high degree nodes (Barabasi, 2016, Chapter 3).

The benchmark introduced in (Lancichinetti et al., 2008) generates networks with power law distributions of community sizes and node degrees, with tunable intra/total ratio (mixing parameter). This benchmark, commonly known by the authors initials (LFR), has been widely accepted and used to test community detection algorithms for static networks. For instance in (Lancichinetti & Fortunato, 2009; Yang, Algesheimer, & Tessone, 2016) a lengthy list of algorithms tested against this benchmark can be found.

3.2.2 Comparing clusterings

A clustering, in our context, is the partition of the set V of nodes of a network into disjoint communities, or formally

$$C = \{c_1, \cdots, c_k\} : \ (c_i \cap c_j = \varnothing \ \ \forall \, (1 \leq i, j \leq k \ \wedge \ i \neq j) \wedge \ \cup_{i=1}^{k} c_i = V \tag{3.1}$$

Comparing communities at successive timesteps is a key requirement to understand community evolution. Comparing clusterings, on the other hand, is critical in our system so that, after all the information required to construct the network at successive time steps is gathered, we can flow nodes resulting in the closest shared information distance between clusterings. Comparing clusterings is an open problem as there is no standard way of measuring the distance between them. Popular methods include several variations of node counting (like the Rand Index) and measures from information theory, like the normalized mutual information and variation of information (VI). A good survey of different methods can be found in (Wagner & Wagner, 2007). We have selected VI, given its robustness, low computational complexity and the fact that it is a true metric (Meilă, 2007).

3.2.3 Temporal community graph generators

There have been some proposals to generate synthetic temporal networks. Most of these generators have as a major goal the benchmarking of temporal community detection algorithms. This is a purpose partially shared by our system, although we are not, in this document, directly discussing what constitutes a temporal community. In spite of recent progresses, community detection in temporal networks is not as mature as in static networks and much work remains to be done. An extensive survey of current methods can be found in (Rossetti & Cazabet, 2018), including a tree-based schema to classify them.

Syntgen can be used as a benchmark generator when control over specific network topologies and structural features are sought, such as community structure, temporal change patterns or assortative mixing. In contrast to other systems, it currently does not support the generation of weighted networks, directed networks or overlapping communities. It is also not the appropriate tool if fine control of node activity is required, such as when a need exists to control link persistence, or to control individual community attributes. Currently, Syntgen does not support

the explicit introduction of change points although these can be induced by varying substantially its input parameters when needed. Beyond benchmarking, we see Syntgen being used to experimentally observe network behavior and we include various examples in section 3.4. The total flexibility to specify community size and node degree sequences makes it possible to load data from an empiric network and analyze the results when other structural parameters are changed.

Other published generators include (Granell, Darst, Arenas, Fortunato, & Gómez, 2015) where the authors propose a generator for simple networks with a cyclic nature based on a variation of the stochastic block model. In (Greene, 2010) the authors have adapted the LFR benchmark (Lancichinetti et al., 2008), while introducing over time ad-hoc modifications to the network. In (Rossetti, 2017) the authors propose RDyn, a system to generate temporal networks respecting a power-law distribution of community sizes and node degrees with tunable clustering and injected lifecycle events that, while disrupting cluster quality, are subsequently re-balanced through re-wiring of node links. In (Bazzi, Jeub, Arenas, Howison, & Porter, 2020) a method is proposed that can represent time evolution as a multilayer network, where each layer corresponds to a network observation (or summary of observations). In contrast with Syntgen, which is a pure temporal network generator, this method can be used for non temporal aspects where each layer corresponds to a different network aspect, such as in the case of multiplex networks where nodes can be part of multiple domains. Authors in (Sengupta, Hamann, & Wagner, 2017) propose a benchmark generator for overlapping communities in temporal networks. They use an extension of the LFR (Lancichinetti et al., 2008) benchmark to support overlapping communities (Chykhradze et al., 2014) and adapt it to the temporal domain. Their method is restricted to power-law distribution of community sizes and node community membership. Intra community links are randomly generated according to some user specified probability. Temporal evolution is also induced by user specified probabilities of community lifecycle events. An obvious difference to our system is the support for overlapping communities. Other differences include an explicit injection of community events while in Syntgen community evolution is a consequence of changes in community size and node degree sequences as time evolves. In Syntgen community lifecycle events can be determined a-posteriori, although an axiomatic system is required for such a task.

All of these generators have obvious affinity with Syntgen. The new contributions introduced by Syntgen include:

- acceptance of granular specifications of community size and node degree sequences which are tested for representation as a simple graph by an original method

- network temporal evolution that minimizes clustering changes

- support for joint distribution of node degrees and node degree temporal correlation

3.3 Syntgen: Description, challenges and contributions

Syntgen is a system to create temporal networks exhibiting community structure that changes over time. It is parametric and modular. The major modules are:

- User specifications. These fall into two separate categories: network topology and heuristics execution.
- Node degree and community size sequence generators. The system includes functions that sample parametric distributions for community size, intra and total node degree, but, as long as they are realizable, any sequences can be provided.
- Network module. Deals with all aspects of network creation, including node to community assignment, degree to node assignment and node to link assignment.
- Transition module. This module manages all aspects of temporal evolution, including heuristics for node flow between timesteps and community lifecycle determination.
- Output module. This module generates all output, both textual as well as machine readable for further analysis. In Table 3.2 we include a summary of all information generated.

Table 3.2. Textual Output of Syntgen

Content	Description
Contingency Matrix	Contingency matrix of communities across time steps
Assortativity Coefficient	Joint node degree distribution
Temporal Degree Correlation	Average Pearsons correlation index for the whole network
Variation of information	VI between clusterings across successive time steps

In the remainder of this section we present the basic algorithmic logic of Syntgen in 3.3.1, the challenges and solutions of building a static network according to user specifications in 3.3.2 and the problem of finding a node flow across time steps that maximizes clustering similarity in 3.3.3.

3.3.1 Syntgen basic logic

The general flow of Syntgen is a sequence of looping steps that produce network snapshots as time progresses. It basically follows algorithm 1.

Syntgen requires from the user at each time step the following graph invariants and parameters:

- a multiset of k positive integers $S = \{s_1, \cdots, s_k\}$, representing a sequence of community sizes

- a bijection of total and intra community degree sequences:
  - an n-tuple of positive integers $D = \{d_1, \cdots, d_n\}$ representing a sequence of node total degrees with $n = \sum_{i=1}^{k} s_i \wedge \sum_{i=1}^{n} d_i \in \{2n : n \in \mathbb{N}\}$
  - an n-tuple of positive integers $E = \{e_1, \cdots, e_n\}$ representing a sequence of node intra-community degrees with $e_i \leq d_i : \ 1 \leq \forall i \leq n$ and $\sum_{i=1}^{n} e_i \in \{2n : n \in \mathbb{N}\}$
- specifications for joint degree distribution and node degree correlation over time
- optionally, a set of nodes O to kill at a step boundary

The user can loosely control the dynamics of the network by changing S, D, E and O at each time step boundary. Depending on the sign of $\sum S_t - \sum S_{t+1} - |O|$ new nodes are implicitly born or additional nodes randomly killed. Correlation and joint distribution parameters have an impact on the wiring of the network. As the data per timestep is gathered, Syntgen executes the following actions:

- A bootstrap static network is built. The input elements are independent (with the exception of the number of nodes and the sum of community sizes, which must match) and it is up to the system to assign links and nodes to communities. We provide parameter-based examples of functions that generate sequences which have been observed in empirical networks (Clauset, Shalizi, & Newman, 2009; M. E. J. Newman et al., 2000; Palla, Barabási, & Vicsek, 2007), sampled from discretized power laws, discretized exponential and binomial distributions, but the user is free to provide his own as adequate to their problem domain.

  Degree assortativity, a topology attribute that varies with the type of network (typically assortative for social, while dissortative for biological networks (M. E. Newman, 2003)), is also parameter driven allowing the user to request a random, weighted assortative or weighted dissortative network.

  To construct the network we use a modified version of the configuration model (Clauset, 2013) in a similar approach to what is found in a popular benchmark for community detection in static networks (Lancichinetti et al., 2008), but developed independently and extended to support joint node degree distributions as described in 3.3.2.5.

  Obviously, not all input specifications are possible and we verify feasibility before generating the network. The problem of whether a given node degree distribution can be expressed as a graph has been covered extensively in the literature (Choudum, 1986; Erdös & Gallai, 1960; Stanton & Pinar, 2011; Tripathi, Venugopalan, & West, 2010), and theorems, like the Erdös-Gallai condition, can be expressed as an algorithm to test graph feasibility. However, with node degrees as tuples of inter and intra-cluster degrees, different conditions apply. We extended the Erdös-Gallai condition to address this problem, developing the corresponding algorithm to halt (or request new input for) the network generation in case input specifications are infeasible.

- After generating the bootstrap network (T_0 network) a T_1 network is generated, again according to user specifications. The user may select a different degree or community size sequences as well as make changes to the network at the end of T_0 (selecting nodes for deletion), according to the requirements of the temporal network to be generated. An additional parameter provides the user with the option of enforcing node degree correlation across timesteps.

  The system then tries to find the closest possible clusterings between successive timesteps. We found the problem to be intractable and impossible to complete in a reasonable amount of time beyond a very small number of communities. To address the inherent complexity, we developed a heuristic based on a greedy anytime algorithm with taboo to search for a solution in an appropriate solution subspace. The objective is to reduce the amount of change (noise) to a minimum, reducing the necessary impact on user specifications. The solution will determine the flow of nodes between communities in timesteps T_0 and T_1.

- This process is repeated for a user-specified number (n) of time steps, evolving the network over a period from T_0 to T_n.

- At each time step the contingency matrix of node/community evolution is produced, and, at the end, the temporal network is created in a machine readable format for further analysis and visualization.

Algorithm 1 General flow of Syntgen. Steps 2,3,6,10 are implemented in the "Network" module. $Build\ Communities$ assigns degrees to nodes and nodes to communities, while $Build\ Network$ does all the network wiring. Steps 7-9 are implemented in module "Transition". The time complexity of Syntgen is determined by the node wiring and transition phases. A complexity analysis is provided in their corresponding sections (3.3.2.5, 3.3.3)

1: $Community\ Size\ Sequence,\ Node\ Degree\ Sequence \leftarrow Sequences\ from\ User$
2: $Build\ Communities\ @\ T_n \leftarrow Community\ Size\ Sequence,\ Node\ Degree\ Sequence$
3: $Build\ Network@T_n \leftarrow Communities$
4: **while** $Remaining\ TimeSteps \neq 0$ **do**
5: $\quad Community\ Size\ Sequence,\ Node\ Degree\ Sequence \leftarrow Sequences\ from\ User$
6: $\quad Build\ Communities\ @\ T_{n+1} \leftarrow Community\ Size\ \&\ Node\ Degree\ Sequences$
7: $\quad Network\ @\ T_n \leftarrow user\ Events$
8: $\quad Flow\ Nodes\ from\ T_n\ to\ T_{n+1} \leftarrow Search\ Most\ Similar\ Transition$
9: $\quad Build\ Network\ @\ T_{n+1}$
10: $\quad Report\ Data\ for\ T_n\ to\ T_{n+1}$
11: $\quad Network\ @\ T_n \leftarrow Network\ @\ T_{n+1}$
12: $\quad TimeSteps \leftarrow TimeSteps - 1$
13: **end while**
14: $Output\ Temporal\ Network$

Syntgen outputs textual information as the network is created overtime, including network metrics, network events and other supporting information. Syntgen also produces the full tem-

poral network in machine readable format that can be input directly to the Gephi (Bastian, Heymann, & Jacomy, 2009) visualization tool.

3.3.2 Creating a static network

Creating a T_0 static network involves the following steps:

- Receiving community size and node degree sequences from the user
- Testing for graphic sequences and requesting new ones if non-graphic
- Randomly assigning nodes without substitution to communities from the bijection of intra (E) and total degrees (D) with $e \in E : e < |c|$
- Wiring nodes using a modified version of the configuration model both for intra links as well as inter links respecting assortative specifications

3.3.2.1 Community, node sequences

Syntgen does not impose specific restrictions on the user input sequences beyond a coherent total number of nodes, and node intra community degrees that are less or equal to their respective total degree. It follows that Syntgen does not enforce community structure per se. The user must provide a ratio of intra to total degree that is conducive to community structure if a clustered network is preferred.

3.3.2.2 Supplied distribution samplers

The user may opt to generate community and node sequences resorting to functionality provided by Syntgen. There are independent and identically distributed (I.I.D) samplers of uniform, exponential and power law distributions. All of our supplied samples of sequence generators accept a ratio (r) of intra to total degree similar to the *mixing parameter* in the LFR benchmark (Lancichinetti et al., 2008). To alleviate rounding artifacts that are more pronounced for nodes with low degree, we employ stochastic rounding instead of rounding to the nearest integer. The authors in (Lancichinetti et al., 2008) point out that allowing the ratio to change can lead to communities containing nodes that have a higher inter than intra degree (due to random fluctuations), but depending on usage, having a fixed intra to total degree ratio may be too restrictive on the desirable network topologies. Therefore, we let the user choose between a fixed $0 \leq r \leq 1$ or Bernoulli distributed ratio with $\mathbb{P} = r$.

Although we do not challenge if specifications as provided by the user or generated by the supplied distribution samplers are conducive to community structure, we do test for disconnected components inside communities by computing the algebraic multiplicity of the zero eigenvalue for the Laplacian of the adjacency matrix of the community. If higher than one, we warn the user, giving the option to continue or abort the network generation.

3.3.2.3 Testing for graphic sequences

To test if user specifications are graphic, that is, if they can be represented as a simple network, we make use of the Erdös-Gallai condition (Erdös & Gallai, 1960) that states that a degree sequence D is graphic if:

$$\sum_{i=1}^{|D|} d_i \in \{2n : n \in \mathbb{N}\} \ \wedge \ \sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{|D|} \min(d_i, k) \ \forall \ (1 \leq k \leq |D|) \tag{3.2}$$

where d is degree and $|D|$ the total number of nodes. We apply 3.2 to the sequence of total degrees and then to every single community using only E, the nodes intra degrees. If completed successfully, we move on to test if the inter degrees sequence F is graphic. For this we reduce the network to a multi-graph where each community becomes a single node and the multi-links are the aggregate inter community links of the base network. It is obvious that $max(F) \leq \sum_{i=1}^{|F|} d_i - max(F)$ is a necessary condition for the graphic property, as otherwise there would be not enough links to satisfy the requirements of the largest community inter degree. But it is also not hard to see that if the total number of inter links is even, the condition above is not only necessary but also sufficient, or formally:

$$\sum_{i=1}^{|F|} f_i \in \{2n : n \in \mathbb{N}\} \ \wedge \ max(F) \leq \sum_{i=1}^{|F|} f_i - max(F) \tag{3.3}$$

To see why, consider a reduced network with 3 nodes (communities), c_1, c_2, c_3 and their respective inter node degree aggregation f_1, f_2, f_3, with $f_1 \geq f_2 \geq f_3$. If $f_1 = f_2 + f_3$, the network is obviously graphic. If $f_1 < f_2 + f_3$ and if $f_1 \in \{2n : n \in \mathbb{N}\}$ then $(f_2 \in \{2n : n \in \mathbb{N}\} \wedge f_3 \in \{2n : n \in \mathbb{N}\}) \vee (f_2 \in \{2n+1 : n \in \mathbb{N}_0\} \wedge f_3 \in \{2n+1 : n \in \mathbb{N}_0\})$ but as $f_1 \geq f_2$ one can always distribute links from c_1 to c_2 and c_3 such that the remainder degrees to be satisfied are equal. If $f_1 \in \{2n+1 : n \in \mathbb{N}_0\}$ then $(f_2 \in 2n+1 : n \in \mathbb{N}_0 \vee f_3 \in 2n+1 : n \in \mathbb{N}_0)$ in which case after one link is added between c_1 and c_2 or c_3 we revert to the previous case.

The above is a proof for a 3 community clustering. To generalize the proof, let's consider the addition of a community to the reduced network resulting in the clustering $C = \{c_1, \cdots, c_4\}$, with node degree aggregation $D = \{f_1, \cdots, f_4\}$, and $f_i \geq f_{i-1} : 2 < \forall i \leq 4$. If we use links from f_1 to satisfy f_4, we get: $f_1 - f_4 \geq f_2 \vee f_1 - f_4 < f_2$. If $f_1 - f_4 \geq f_2$ we reduce to the previous proof as $f_1 - f_4 + f_2 + f_3 \in \{2n : n \in \mathbb{N}\}$ and $f_1 - f_4 \leq f_2 + f_3$ (remember that $f_3 \geq f_4$).

If $f_1 - f_4 < f_2$ then to reduce to the previous 3-community proof we should have $f_2 < f_3 + (f_1 - f_4)$. This is easy to prove by contradiction as $f_2 > f_3 + (f_1 - f_4)$ is impossible, given that $f_1 > f_2$ would force $f_3 - f_4 < 0$ which violates the problem statement. So by contradiction and induction we prove the condition for the graphic property of the inter links part of the network.

In conclusion, a network with $|V|$ nodes and $|C|$ communities with size sequence S, each

with a bijection of intra and inter degree sequences respectively E_{c_i}, F_{c_i} $\forall i \in \{1, \cdots, |C|\}$ is graphic under the condition in equation 3.4.

$$\forall i \in \{1, \cdots, |C|\} : \sum_{j=1}^{s_i} e_j^{c_i} \in \{2n : n \in \mathbb{N}\} \ \wedge \ (\sum_{j=1}^{k} e_j^{c_i} \leq k(k-1) + \sum_{j=k+1}^{s_i} \min(e_i, k) \forall (1 \leq k \leq s_i)$$

$$\wedge$$

$$\sum_{i=1}^{|V|} f_i \in \{2n : n \in \mathbb{N}\} \ \wedge \ max(F) \leq \sum_{i=1}^{|V|} f_i - max(F) \tag{3.4}$$

3.3.2.4 Node assignment

Syntgen assigns nodes to communities randomly at time step T_0 from the pool of available nodes, avoiding communities with cardinality smaller than the node intra degree. From T_1 onwards nodes keep their community membership except to honour new community size sequences. The process of minimizing membership changes is covered in section 3.3.3. The user can indirectly control node degree temporal correlation by influencing degree selection from the supplied total degree sequence thru the shape parameters of a beta distribution used to sample the ordered sequence. When $\alpha = \beta = 1$ it reverts back to the uniform distribution.

3.3.2.5 Configuration model

In Syntgen we based the generation of networks with user specified degree distributions on a modified version of the configuration model (CM) (Barabasi, 2016, Chapter 3). We use this modified version to wire nodes inside communities (one community at a time, as if they were separate networks) and to create inter community links.

The CM can create a network based on arbitrary sequences D of node degrees. To this end, it expands the sequence into a list of $(\sum D)$ node "stubs" that are randomly paired, creating links. See figure 3.2 for an example.

Figure 3.2. Wiring the configuration model. Example of setting up a first link between node stubs for a network with $n = 10$ nodes, 15 links and degree sequence $D = \{4, 4, 3, 3, 4, 3, 3, 2, 2, 2\}$. Stubs are randomly chosen and as long as $\sum_{i=1}^{n} D_i$ is even, the process always concludes, albeit with multi links and self loops.

For our purpose, the standard CM presents two difficulties. The first is that nothing prevents a stub from linking back to another stub belonging to the same node, or linking the same nodes multiple times, both of which are incompatible with our aim of building a simple network with

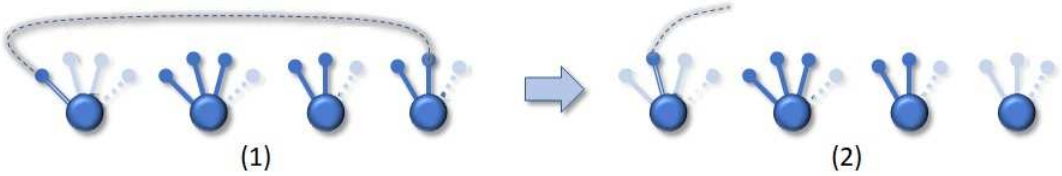
no self loops and no multi-links. The second is that we want to provide the user with some capacity to control joint degree distribution, while the CM results in the following fixed joint distribution:

$$p_{ij} = \frac{k_i k_j}{S - 1} \tag{3.5}$$

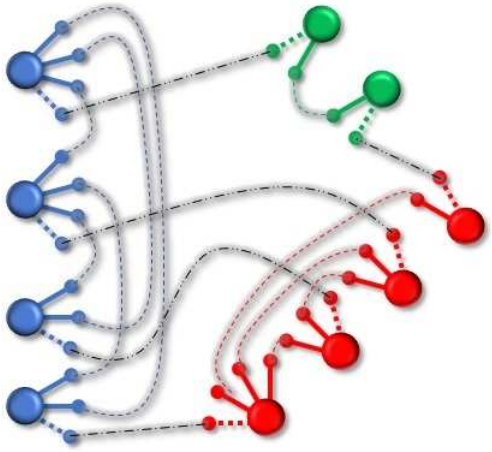
where k_n is the number of stubs of node n and S the total number of stubs = $\sum D$.

As the network grows, the probability of self-loops and multi-links decreases. This probability varies with the actual node degree distribution, but it is not unreasonable to disregard self-loops and multi-links when building the network (see figure 3.3 for an example), considering however that, (1) stub pairing can fail before all stubs are assigned (that is, a node can have unlinked stubs with no candidate stubs remaining), and (2) that equation 3.5 is no longer representative of the degree joint distribution.

(a) We use a modified CM, and apply it to one community at a time for intra-community links (full stroke linestyle) and to the whole network for inter community links (dashed linestyle)



(b) Our modified version of the CM deletes from the link candidate list all the remaining stubs of the linking node (double stroke linestyle) (1), and all the remaining stubs of the linked node for subsequent stubs from the same node (2)



(c) Example of the fully connected network

Figure 3.3. CM Plots of a network 3 communities (Blue, red and green), with community size sequence $\{4, 4, 2\}$ and total and intra degree sequences $\{4, 4, 4, 3, 3, 3, 3, 2, 2, 2\}$, $\{3, 3, 3, 2, 2, 2, 2, 1, 1, 1\}$.

The first problem can be circumvented by selectively rewiring nodes randomly from a pool of candidate nodes (those that could satisfy the outstanding stubs but are otherwise taken else-

where). As we test that the network specifications are graphic beforehand, and the rewiring process is ergodic, this always completes successfully.

The second problem is less relevant as we aim to generate networks with tunable joint degree distribution. We modified the CM so that instead of connecting stubs I.I.D. over a uniform distribution, we connect them I.I.D. over a beta distribution from the ordered node degree sequence. As the probability density function (pdf) increases towards the rightmost side of the distribution domain, correlation increases, and vice-versa. The α and β shape parameters of the Beta distribution are specified by the user and enable flexible pdf shapes. Using these parameters, the user can influence the level of the network correlation, subject to structural cutoffs (Boguñá, Pastor-Satorras, & Vespignani, 2004; M. E. Newman, 2003).

In our CM implementation we visit every other node for each node in the network. When stub pairing fails, additional node "visits" are required, but, as the probability of failure is directly tied to the probability of self-loops and multi-links, and these tend to zero as the network grows, using asymptotic notation we obtain a time complexity of $\mathcal{O}(|V|^2)$, where V is the network node set. Additionally, we may need to sort the list of candidate nodes to honor degree correlation requests, increasing the overall complexity of this phase of the algorithm to $\mathcal{O}(|V|^3 \log |V|)$.

3.3.3 Minimizing Shared Information Distance

Once we have constructed the network N_t at time t, injected user changes, created the network N_{t+1} (all based on user specifications), and created adjustments in communities for dead and new nodes, so that the number of nodes across steps remains the same, all that is left is to flow surviving nodes from one network to the next. We want to perform this node flow in such a way that the clusterings C^t and C^{t+1} are as similar as possible, in this way minimizing the changes beyond the user specifications. To measure the changes, we need a way of comparing clusterings. As mentioned previously, there are several approaches to this problem, from node pair counting to information based distance measures.

There is no best approach as explained in (Meilă, 2007), and we have opted to use the Variation of Information:

$$VI(X;Y) = -\sum_{i=1}^{k}\sum_{j=1}^{l} r_{ij}[log(\frac{r_{ij}}{p_i}) + log(\frac{r_{ij}}{q_j})] \tag{3.6}$$

where $X = \{x_1, \cdots, x_k\}$ and $Y = \{y_1, \cdots, y_l\}$ are clusterings of a given set S, with $n = |S|$, and $r_{ij} = \frac{x_i \cap y_j}{n}$, $p_i = \frac{|x_i|}{n}$ and $q_j = \frac{|y_i|}{n}$. Our choice of VI is based on its algorithmic simplicity and on the fact that it is a true metric (Kraskov, Stögbauer, Andrzejak, & Grassberger, 2005a), respecting positivity, symmetry, and the triangle inequality.

But let's set briefly aside the proposed method of comparing clusterings and consider the search space of feasible node flows between N_t and N_{t+1}. One way of looking at the problem

is to coalesce N_t and N_{t+1} into the weighted bipartite network $G(C_t, C_{t+1}, U)$, where the nodes are the communities at successive timesteps and the U the weighted links representing the node flows between them.

It is easy to see that the search space consists of the solutions to an under-determined system of Diophantine equations $Ax = B$ where A is the incidence matrix of the fully connected bipartite network $G = (C_t, C_{t+1}, L)$ and B is the vector $\{|C_i^t|_{i=1}^k \cup |C_j^{t+1}|_{j=1}^l\}$. As $rank(A) = |B| - 1$, one line of matrix A and the corresponding entry of vector B can be removed. Dimensionality can be further reduced as $\sum[x]$ is known, and thus one element of x can be determined from the others. Every solution in the solution space is a vector x whose elements are the number of nodes (u) that should be transferred from communities in C_t to communities in C_{t+1}.

Formally we want to find $C_{t+1}\ s.t.\ min(VI(C_t, C_{t+1}))$ with $Ax = B$ as defined above.

In topological terms, the space of the solution is a lattice contained in an $n-1$ dimensional polytope, where $n = |C_t| \times |C_{t+1}|$, bounded by $n-1$ positive halfspaces (as $x_i \geq 0 :\ \forall i$), and by $|C_t| + |C_{t+1}| - 1$ hyperplanes defined by the equations in $Ax = B$. The number of solutions is equal to the number of lattice points. Counting lattice points in such a polytope is not an easy task (Loera, 2005) and quickly becomes intractable. Barvinok proposed an algorithm for lattice point counting in (Barvinok & Pommersheim, 1999) that has been implemented in systems like Latte (Baldoni et al., 2014), software that counts lattice points and performs integration inside convex polytopes. Some experiments we ran in Latte that illustrate the size of the problem can be seen in table 3.3.

Clustering @ T	Clustering @ T+1	Number of solutions
$\{20, 16, 12\}$	$\{24, 13, 11\}$	$6.46000E+03$
$\{16, 16, 16\}$	$\{16, 16, 16\}$	$1.17810E+04$
$\{13, 11, 10, 10\}$	$\{14, 12, 9, 9\}$	$7.80605E+06$
$\{1300, 1100, 1000, 1000\}$	$\{1400, 1200, 900, 900\}$	$1.58534E+24$
$\{13, 11, 10, 10, 9\}$	$\{14, 12, 9, 9, 9\}$	$1.09501E+11$
$\{1300, 1100, 1000, 1000, 900\}$	$\{1400, 1200, 900, 900, 900\}$	$3.18145E+41$

Table 3.3. Solution space, as reported by the count function of Latte, for 6 examples of clustering pairs. As can be seen, flowing even a small number of communities generates a search space that is for all purposes intractable.

The solutions that are of interest to us will have a high degree of sparsity as we are looking for similar clusterings and, intuitively (and experimentally), a high dispersion of nodes across communities will not be conducive to similarity. Higher sparsity solutions correspond to surface features of the polytope lattice, that is, in decreasing sparsity order: vertices, edges, ridges, cells, facets and so on, basically the $1, \cdots, n-1$ elements of an n-dimensional polytope. We based our heuristic on this intuition, limiting our search space to the hull of the polytope (see figure 3.4) This can reduce the space significantly depending on the polytope geometry.

To scan the space we find the nullspace of A, formally $ker(A) = \{x \in \mathbb{N}^n : Ax = 0\}$, where $n = |C_t| \times |C_{t+1}|$, and one solution x_i to $Ax = B$. By linearly combining x_i with

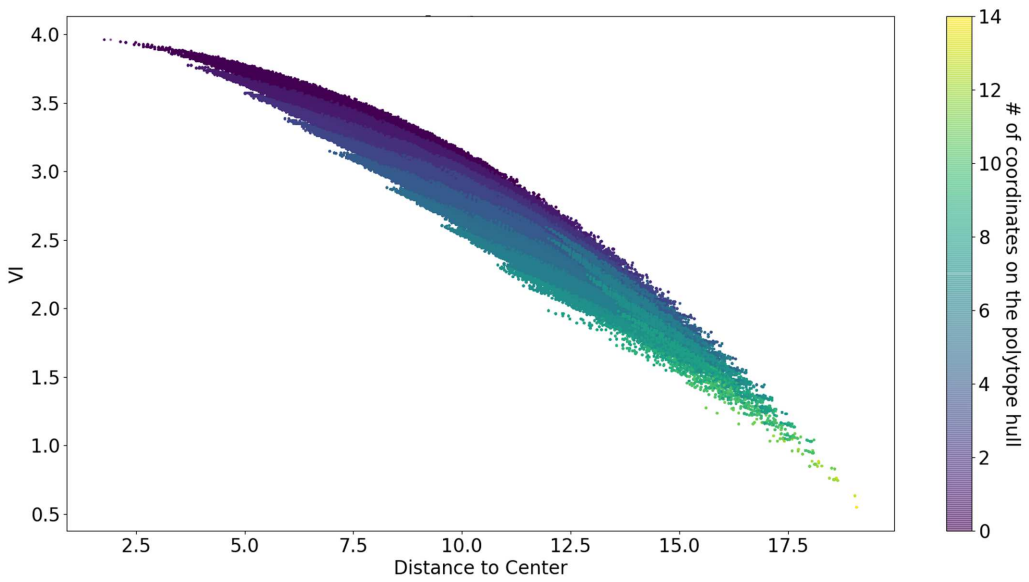


Figure 3.4. Comparing Clusterings similarity as a function of spacial location Plot of all the 16,799,002 possible solutions of flowing a clustering with community size sequence of {13, 13, 12, 10} to {15,11,11,11}. We compare similarity, as measured by the variation of information, against distance to polytope center and number of polytope surface coordinates in the solution vector. The polytope "center" is computed as the vector $\left(\frac{max(x_i)\times n}{\sum_{j=1}^{f} max(x_j)}\right)_{i=1}^{f}$, where the vector x is the number of nodes flowing between communities (the sequence of link weights of the fully connected bi-partite network, see section 3.3.3), n is the total number of nodes, f the number of possible flows and $\max(x)$ the maximum number of nodes flowing between two communities. It is clearly visible that there is a strong positive correlation between these quantities.

$ker(A)$ we can span the set of solutions to $Ax = B$. Finding a single solution is trivial, all that is needed is to flow nodes from C_t to C_{t+1} until no more nodes are left in C_t. Finding an optimal solution is, however, at least as hard as the partition problem, a well documented NP-complete problem (Korf, 1998), that asks if a given multiset of positive integers can be split into two subsets that sum to the same amount. We prove this by reducing the partition problem to our problem, and by applying the method described in (Cormen, Leiserson, Rivest, & Stein, 2009, p.1078, Lemma 34.8). It works this way: consider a general partition problem, with a multiset $M = \{m_1, m_2 \cdots m_k\}$ with $\sum_{i=1}^{k} m_i = n$. This reduces polynomially to a problem where a source clustering $S = \{s_1, s_2 \cdots s_k\}$ has $\{|s_1|, |s_2| \cdots |s_k|\} = M$, and a target clustering $T = \{t_1, t_2\}$, where $|t_1| = |t_2| = \frac{n}{2}$. There is a solution where no source cluster is split in the target clustering, that is when either $s_i \subseteq t_1$ or $s_i \subseteq t_2$, if and only if, there is a solution to the partition problem. It is easy to see and prove that this condition results in the most similar clusterings, as dispersion of nodes, from individual clusters in S across t_1 and t_2, increases the variation of information.

To find a more promising starting point for our heuristic space search, we have implemented a pool of five simple algorithms, with polynomial complexity on the number of communities, that were experimentally best performers among themselves. They implement one-pass greedy

heuristics with an objective function related either to sparsity or similarity. Experimentally, although any one of the five may achieve best performance, one of them clearly outperforms the others as the number of communities increase (see figure 3.5).

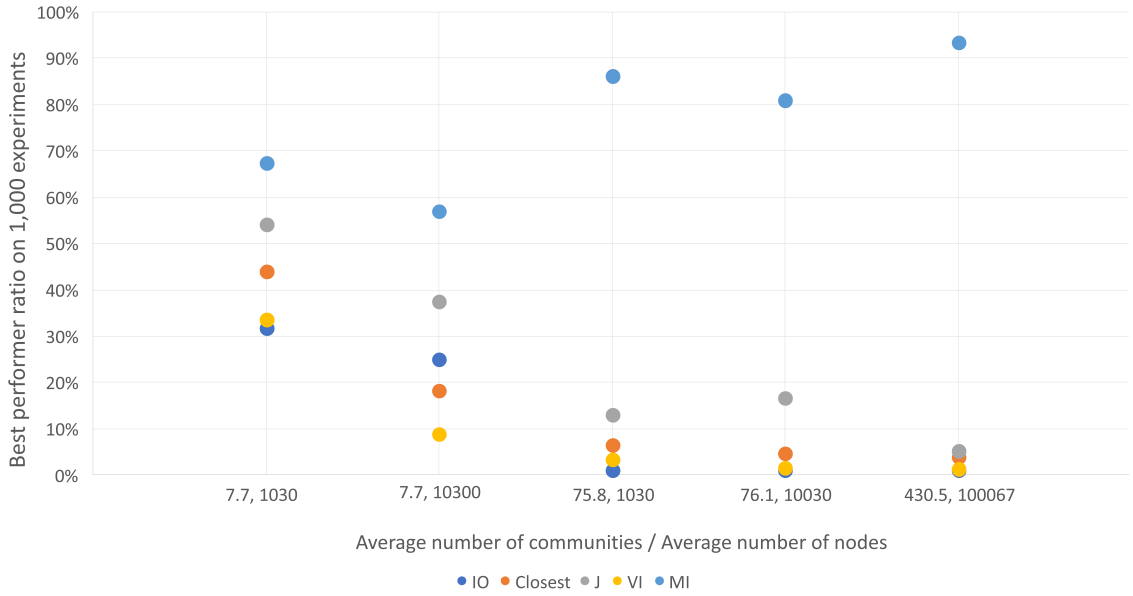


Figure 3.5. Relative performance of a pool of 5 simple algorithms to select a starting point for a space scan All algorithms achieve top VI-based similarity in some of the 1,000 random runs, but one (MI, based on minimizing the increment of mutual information) vastly outperforms all others as the number of communities increases.

In our space scan heuristic, we use these solutions (or the best of them) as starting points for our space search. These simple algorithms are single pass over the successive community sets and run in at most $|C_t| + |C_{t+1}|$ time, where C is the community set, so their asymtoptic complexity is $\mathcal{O}(|C|)$ and can be ignored when compared to the algorithmic complexity of the solution space search. This search is accomplished by an anytime algorithm that greedily scans the solution polytope hull for the lowest VI avoiding previously visited solutions (see algorithm 2). To halt the algorithm, the user can specify thresholds for search restart after a certain number of failed improvement trials and a certain number of failed restarts. Each search elementary step visits two communities from two successive timestep (C_t, C_{t+1}), and all communities are potentially selectable, (although in practice this rarely happens as the output from the simple algorithms is already fairly optimized). This results in a complexity of $\binom{|C_t|}{2} \times \binom{|C_{t+1}|}{2}$ per single search with an overall asymptotic complexity of $\mathcal{O}(|V|^4 \times l \times g)$ if we consider the same number of communities at successive time steps and where l and g are respectively the number of local and global tries. Adding the time complexity for wiring the network (see 3.3.2.5), we get a final complexity of $\mathcal{O}(|V|^4 \times l \times g + |V|^3 \log |V|)$.

Searching the hull of the polytope vastly reduces the search space in most circumstances, but, as the network grows, the probability of improving on the results from the pool of simple algorithms decreases. For very large networks the user may select to proceed with the best

Algorithm 2 Anytime greedy algorithm with taboo

$currBest \leftarrow min(Solution(SimpleAlgorithmPool))$
$bestVI \leftarrow VI(currBest)$
$globalTries \leftarrow 0$
$visited \leftarrow \emptyset$
while $globalTries \leq globalTriesThreshold$ **do**
    $localTries \leftarrow 0$
    $globalTries \leftarrow globalTries + 1$
    while $localTries \leq localTriesThreshold$ **do**
        $localTries \leftarrow localTries + 1$
        $localBest \leftarrow MAXFLOAT$
        for all $v = vector \in ker(A)$ **do**
            $n \leftarrow ((na \times v + currBest \in solutionSpace) \wedge ((n+1) \times v + currBest \notin solutionSpace)$
            $newSol \leftarrow n \times v + currBest$
            if $newSol \notin visited$ **then**
                if $VI(newSol) \leq localBest$ **then**
                $localBest \leftarrow VI(newSol)$
                $newSolLocal \leftarrow newSol$
                end if
            end if
            $n \leftarrow ((n \times v + currBest \in solutionSpace) \wedge ((n-1) \times v + currBest \notin solutionSpace)$
            $newSol \leftarrow n \times v + currBest$
            if $newSol \notin visited$ **then**
                if $VI(newSol) \leq localBest$ **then**
                $localBest \leftarrow VI(newSol)$
                $newSolLocal \leftarrow newSol$
                end if
            end if
        end for
        if $localbest = MAXFLOAT$ **then**
            Break Global Tries (Dead end)
        else
            $visited \leftarrow visited \cup newSolLocal$
            if $VI(newSolLocal) \geq bestVI$ **then**
                $localTries \leftarrow localTries + 1$
            else
                $bestVI \leftarrow VI(newSolLocal)$
                $currBest \leftarrow newSolLocal$
                $localTries \leftarrow globalTries \leftarrow 0$
            end if
        end if
    end while
end while

result from the pool and forego the heuristic search for the sake of expediency.

In figure 3.6 an example of an exhaustive search of a very simple temporal network with a total of 279 solutions can be found to illustrate the method.

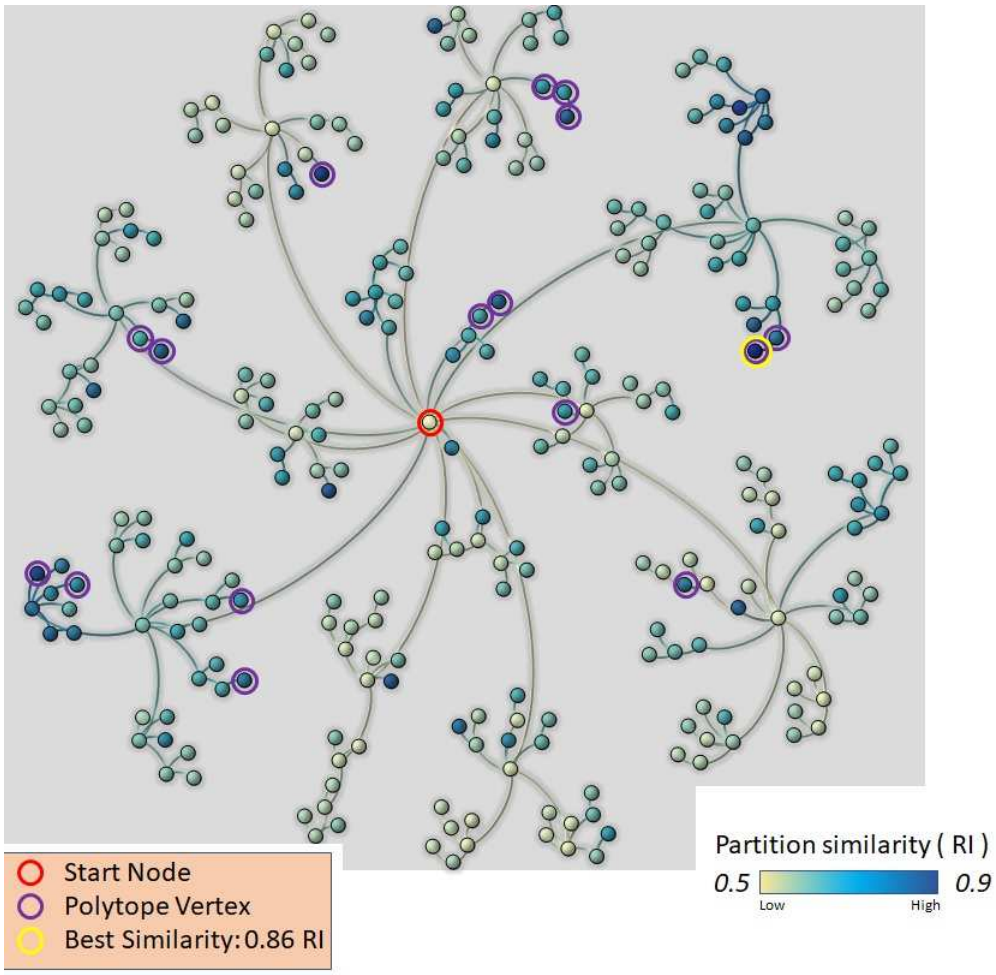


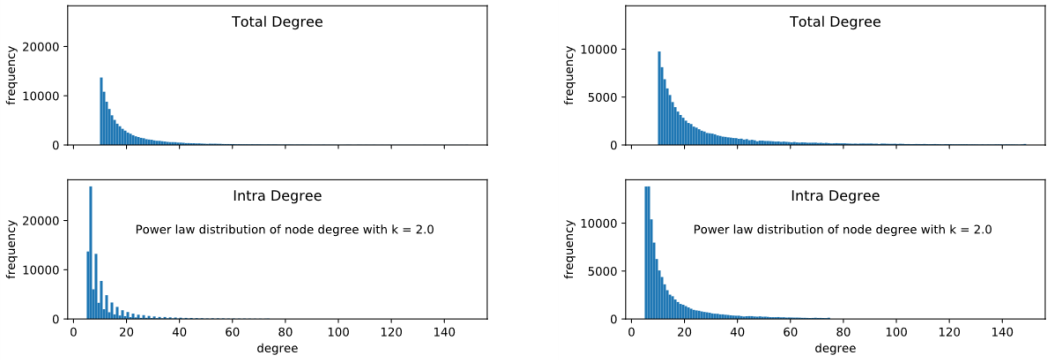
Figure 3.6. Example of a heuristic search to minimize information distance. Map of a full search of a small network transition from $\{10, 8, 6\}$ to $\{12, 10, 2\}$. Note that vertex points of the solution lattice have higher than average similarity as measured by the Rand Index.

3.4 Experiments

Syntgen can generate many variants of temporal networks with community structure respecting user defined distributions of community sizes and node degree. In this section we present and discuss network metrics from generated networks according to varying input parameters. Given the time-slicing nature of Syntgen, some of the experiments highlight results from a single point in time, with the understanding that input parameters can be changed by the user on every time slice.

Sample Distributions

Syntgen provides distribution samplers of node degrees and community sizes. In figure 3.7 we can see an example of a power-law degree distribution and the effects of different rounding approaches.



(a) Rounding to nearest integer **(b)** Stochastic rounding

Figure 3.7. Node degree distribution of networks generated with 100,000 nodes, mix ratio of 0.7, and total node degree varying from 10 to 150. The artifact minimizing effect of stochastic rounding can clearly be seen in these examples.

Mix Ratio

We studied experimentally the impact of using a fixed vs Bernoulli distributed mix ratio (μ). As expected we did not observe significant differences between both approaches when run over 11 time steps, as can be seen on figure 3.8.

Joint Degree Distribution

As discussed in section 3.3.2.5, assortativity is tunable by the user thru the shape parameters of the Beta distribution, affecting link generation. However, dependent on network structure, it may be impossible to generate a network with positive correlation. In figure 3.9 we plot all links of a network with 10,000 nodes and power-law distribution of node degrees and community size on a two-dimensional graph showing the degrees of their connected nodes. As the network is non directional the plots are symmetrical when reflected about the diagonal. The result of applying shape parameters to influence correlation can be clearly seen. We confirm previous findings, noting additionally that clustering has a strong influence on correlation behaviour, potentially limiting the possibility of generating a positively degree correlated network.

It is intuitive that a highly assortative network increases the tendency for groups of nodes to ”clump” together, which suggests that clustering quality may be affected in highly assortative networks even when the user specifies intra and total degree bijections that are conductive to community structure. To validate this intuition experimentally, we built temporal networks with

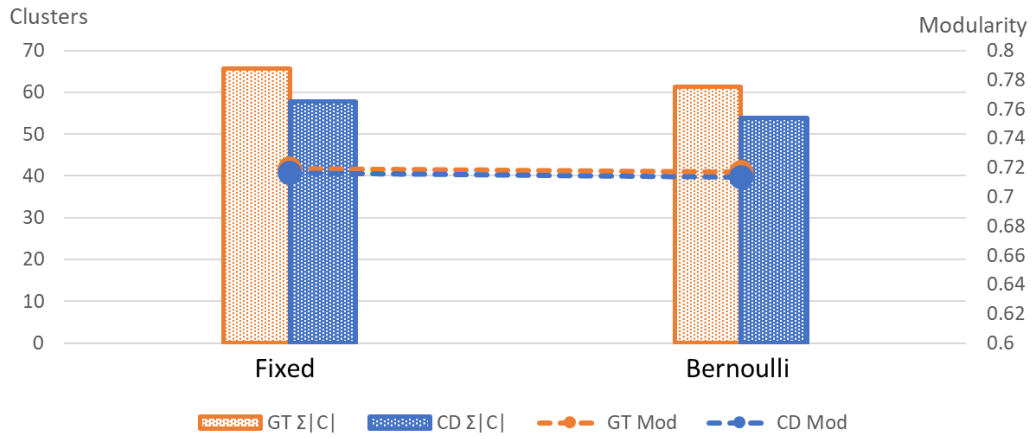


Figure 3.8. Fixed versus Bernoulli distribute mix ratio Two networks averaged over 11 time steps, with 10,000 nodes, mix ratio $\mu = .7$, power law distribution of community size ($K_c = 1.5$) and node degree ($K_n = 2.5$), displaying ground truth modularity and modularity as computed by the community multilevel algorithm (Blondel et al., 2008). Differences in modularity between experiments are negligible. The differences in number of communities found between the ground truth and the community detection algorithm can be attributed to the resolution limits of the algorithm used for detection (Fortunato & Barthélemy, 2007).

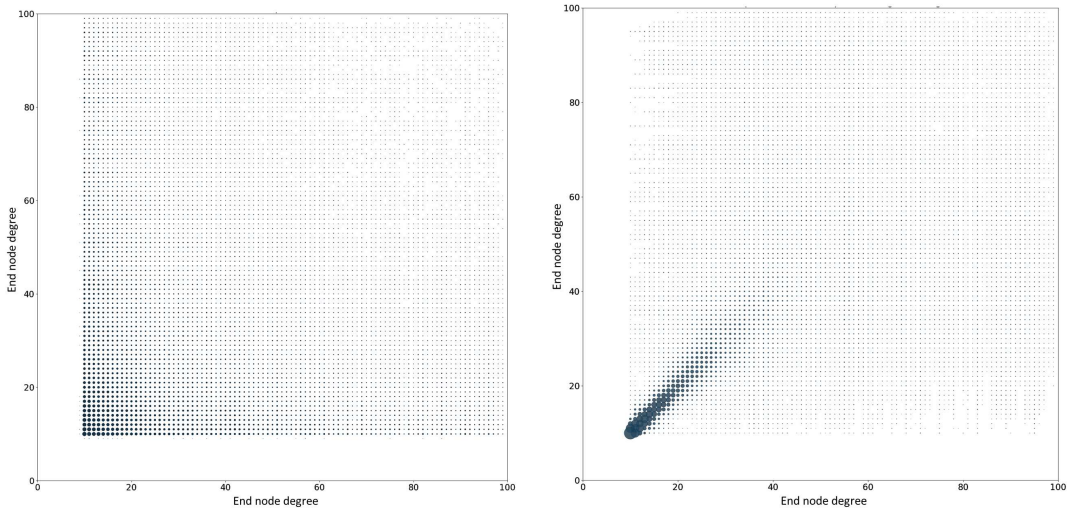
the same community size and node degree sequences generated from powerlaw distributions, at varying assortative indexes. The networks had approximately 10,000 nodes and 60 communities, intra to total link ratio of 0.7 and density of 0.2% and were left to evolve over 100 time steps. Determining clustering quality is sometimes problematic (Dao et al., 2017). For instance, in networks generated by Syntgen, all being equal except for joint degree distribution, there should not be any difference in the ground truth clustering modularity or community conductance. On the contrary, the average minimum cut and average global clustering coefficient for all communities, and the global clustering coefficient for the whole network, are metrics that will be affected by the network assortativity. Results can be seen in figure 3.10. The observed reduction in the minimum cut and the increase in the clustering coefficient inside the communities, as the network becomes progressively more assortative, seems to confirm the intuition.

Temporal Correlation

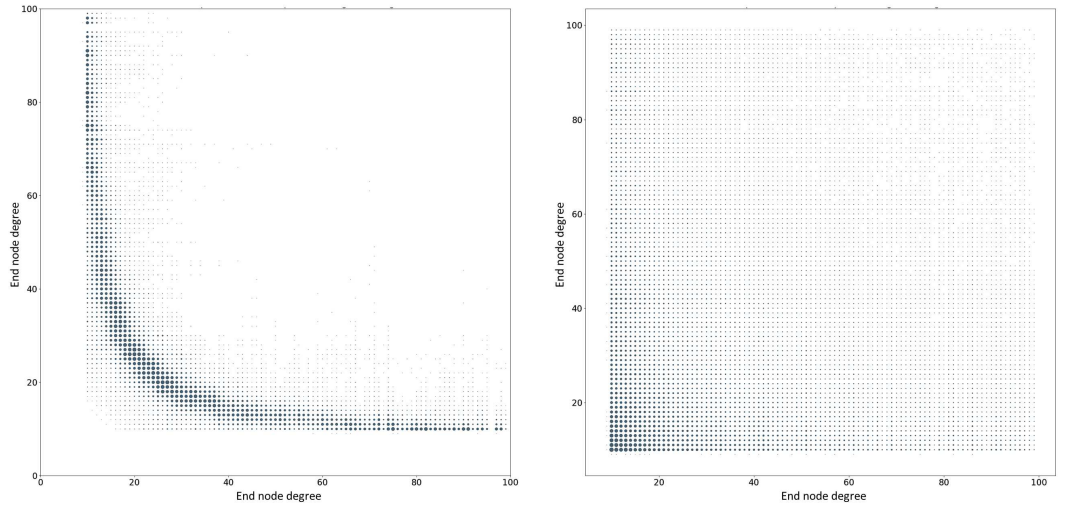
We use the same technique of sampling a beta distribution to influence the evolution of node degree. The user can change the distribution shape parameters to sustain a temporally homogeneous node degree, or to generate nodes that are cyclically active. Figure 3.11 shows the impact of varying shape parameters on the temporal node degree correlation.

Syntgen does not directly provide a facility to control link persistence over time. Obviously network metrics, such as density, have a self-evident impact on edge persistence. Other metrics may have a less obvious influence. We experimentally tested how edge persistence varied with degree assortativity and degree temporal correlation.

Using the same network as described above, we did not find any significant correlation be-



(a) Uncorrelated network ($\alpha = 1, \beta = 1, \overline{D} = 20$) **(b)** Assortative network ($\alpha = 21, \beta = 1, \overline{D} = 20$)



(c) Dissortative network ($\alpha = 1, \beta = 21, \overline{D} = 20$) **(d)** Failed assortative network ($\alpha = 21, \beta = 1, \overline{D} = 25$)

Figure 3.9. Assortative Experiments on power-law networks with 10,000 nodes, varying the average degree from 20 to 25 and maximum degree from 100 to 500, with an average community size of ≈ 168. Every point on the chart is a link with (x, y) coordinates representing the connecting nodes degrees. The point size is directly proportional to the total number of links with equal coordinates. As can be seen, as we increase the average node degree from 20 to 25 by increasing maximum node degree from 100 to 500, it is no longer possible to generate a correlated network with stated metrics even with aggressive beta distribution shape parameters.

tween edge persistence and degree assortativity (average 5.6%, minimum 5.5%, maximum 5.7% in an assortativity index range of $[-0.67, 0.59]$) with all other parameters invariant. However node degree temporal correlation does seem to have a positive correlation with edge persistence in non-assortative networks as can be seen in figure 3.12. In a highly assortative network (0.59 assortative index) the correlation is more complex, with node persistence increasing as the network moves away from non temporally correlated node degrees which we believe warrants further study.

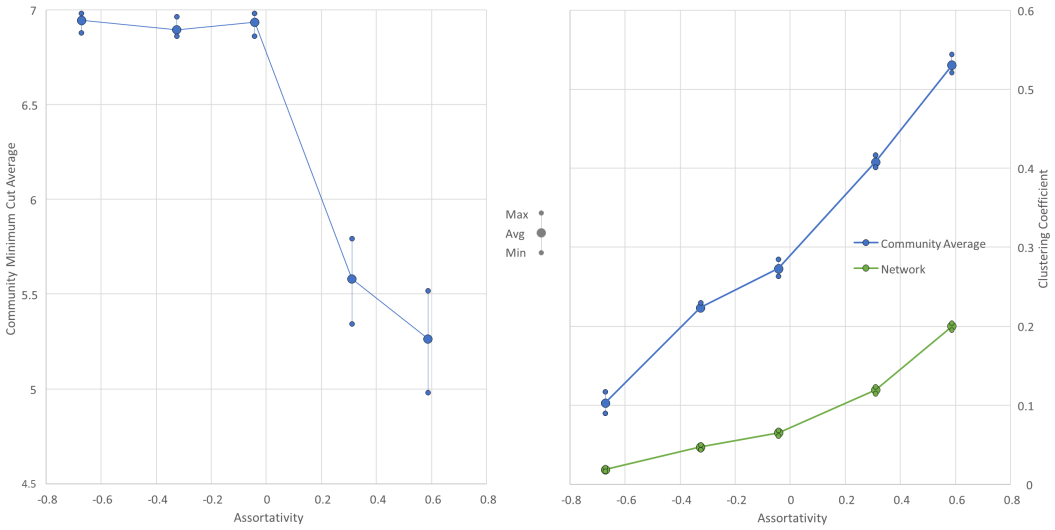


Figure 3.10. Minimum cut and global clustering coefficient as a function of the network assortativity index Average, maximum and minimum of the averages of 100 observations of the evolution of 5 networks under varying assortativity indexes.

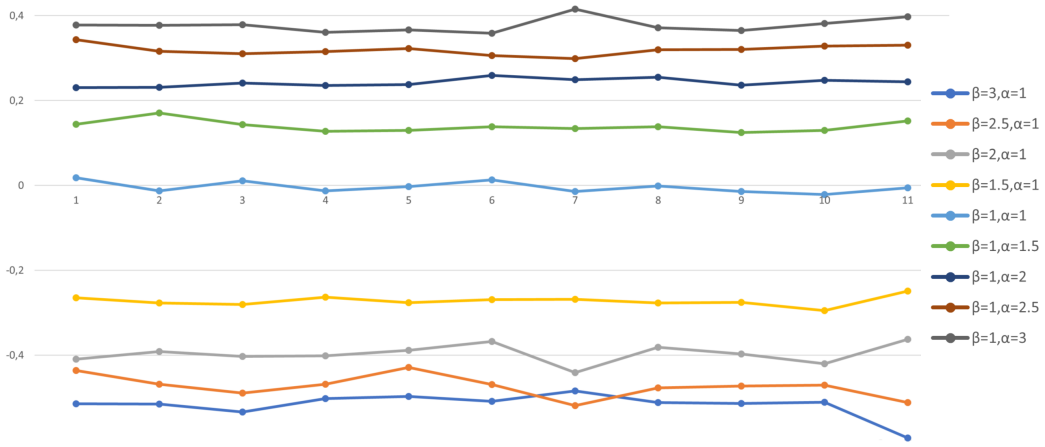


Figure 3.11. Temporal node degree correlation as a function of the Beta Distribution shape parameters Evolution of the node degree Pearson's correlation at 11 successive time steps for different specifications of the Beta distribution shape parameters.

Sample Network

Currently Syntgen outputs machine readable networks in CSV format adequate for loading into Gephi (Bastian et al., 2009). In figure 3.13 an example of a Syntgen generated temporal network with lifecycle events can be seen. A simple analysis of the transition from T_2 to T_3 can be found in table 3.4. The events were categorized as a function of the Jaccard Index (Jaccard, 1912) between communities, based on an external threshold to indicate community continuation.

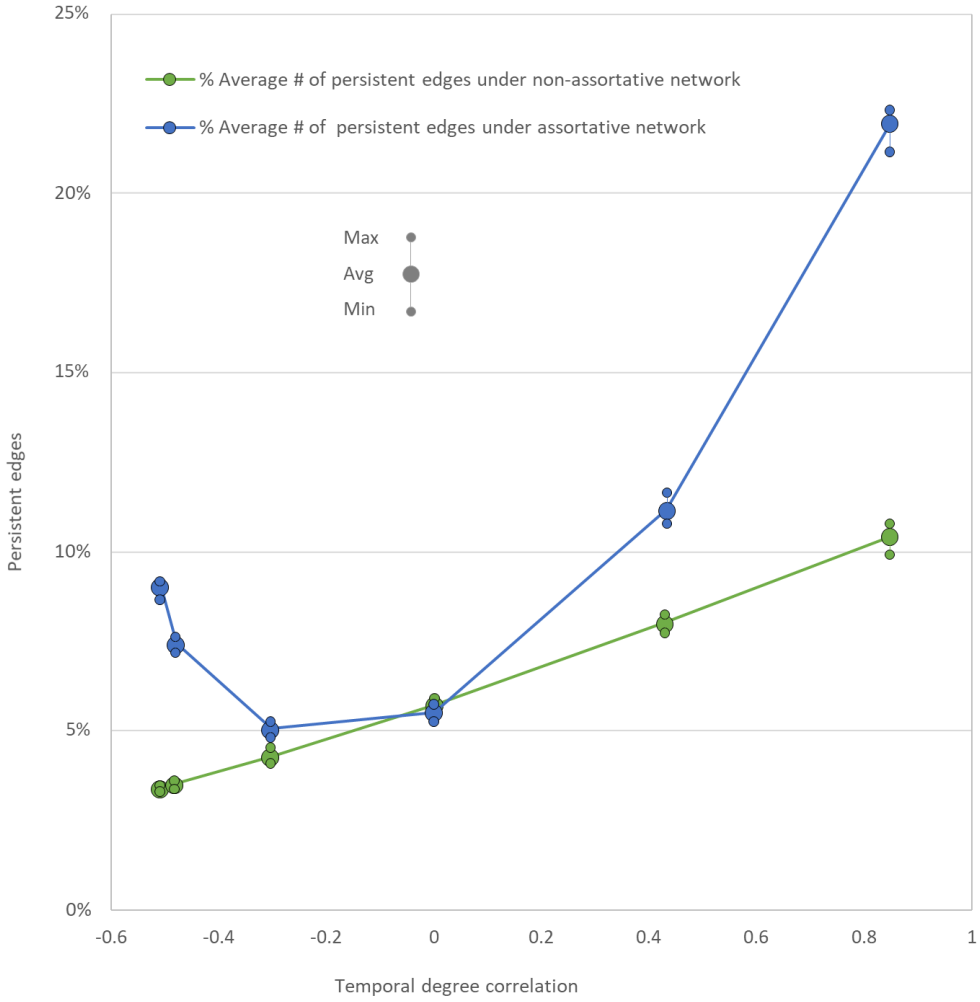


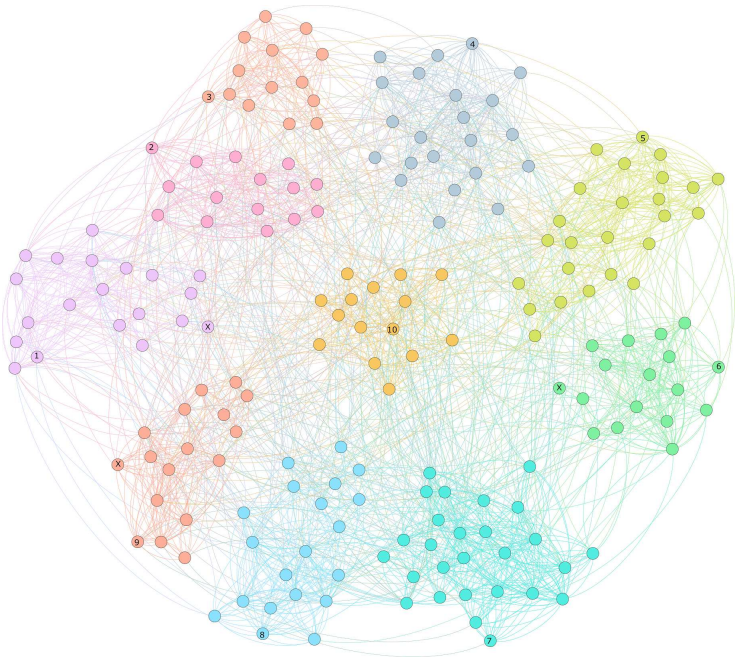
Figure 3.12. Edge persistence as a function of node degree temporal correlation on non-assortative and highly assortative networks Average, maximum and minimum of averages of 100 observation of the evolution of 5 networks under varying node degree temporal correlation.

3.5 Remarks, Discussion and Conclusion

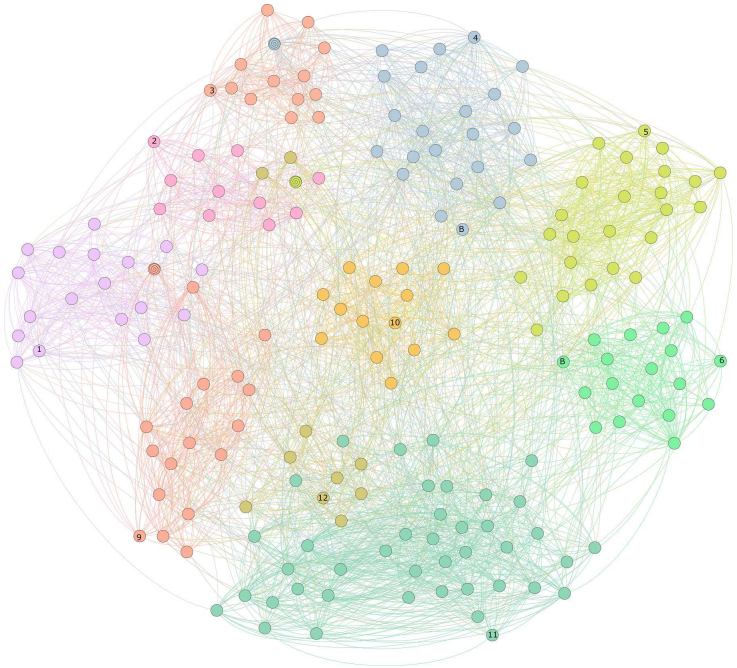
Syntgen is a network generator with constraints. Links between nodes are created I.I.D. over explicit and implicit user specifications. Although a single instance may deviate from the required specifications, the average of a set of generations will converge asymptotically to those specifications.

When using the supplied distribution samplers for node degree and community size, the size of the network has an impact on how closely specifications can be followed. For example, a network with a low average number of communities, will have community size distributions that are likely not recognizable when compared to parameter derived expectations.

It is also of note that, although Syntgen proceeds basically I.I.D. when wiring the network, every time a ”dead-end” is encountered on graphic specifications, the affected process is restarted after re-wiring adjustments. For instance, if, while generating intra links inside a



(a) Network at time T_2 with 10 communities, numerically and color identified.



(b) Network at time T_3 with 9 communities with merge and split events.

Figure 3.13. Nodes marked "X" are examples of nodes to be killed across the transition to T_3. Nodes with a spiral are example of nodes that changed communities, and nodes marked "B" are examples of new born nodes. Community 8 split from T_2 to T_3 into community 12 and split-merged with community 5 to community 11.

community, a node exhausts its list of candidate nodes before satisfying its degree, other established links will be broken so that the process can proceed to satisfaction. This, in practice, "breaks" the I.I.D. aspect of the system, although for most specifications the impact will be marginal depending on the density of the communities and of the network.

Community	Event @ end of time T_2
3	Continues shrinking in 3
10	Continues shrinking in 10
2	Continues shrinking in 2
6	Continues growing in 6
9	Continues growing in 9
8	Split into [12, 11]
8	Merged Into 11
1	Continues shrinking in 1
4	Continues growing in 4
5	Continues in 5
7	Merged Into 11
Community	Event @ beginning of time T_3
12	from Split 8
2	Continued shrinking from 2
3	Continued shrinking from 2
10	Continued shrinking from 10
9	Continued growing from 9
6	Continued growing from 8
1	Continued shrinking from 1
4	Continued growing from 4
5	Continued from 5
11	from Split 8
11	Merged from [8, 7]

Table 3.4. Community events on time step transition
Note community 8 as it splits into 12 and merges into 11 continuing in both communities

Joint degree distribution specifications and node affinity over time is influenced by user parameters, but is also restricted by network wiring requirements and structural cutoffs (Dorogovtsev & Mendes, 2002). This is the reason why it is not possible to directly specify node joint degrees or temporal correlation.

The ratio of intra to total node degree has a direct impact on the clustering modularity at any given time. The only node information kept by Syntgen is the intra-community and inter-community node degree that the user provides, and its linked nodes and community membership, generated by Syntgen. If specifications are not conducive to a clustered network, community membership will not be recovered from the network structure.

In fact, modularity is affected not only by the above ratio, but also by assortativity specifications. In a highly correlated network it is possible that the clustering generated by Syntgen does not exhibit maximum modularity, which can be verified experimentally. The intuition is that, as nodes exhibit connection preferences, communities within communities may appear, resulting in improved modularity with a larger number of communities.

The main aim of Syntgen is to provide researchers in network science a tool flexible enough to generate temporal networks that approximates the topology observed in empirical networks. Syntgen can help where real data is not easily accessible, but whose structure and topology

is known. In the process of building Syntgen, we developed a method to determine if intra and inter node degree and community size sequences are graphic, and a heuristic to find the node flow that results in the closest clusterings at successive time steps, given a network and a community size sequence.

We plan to use the developed search heuristic to determine change points in temporal networks with community structure. The intuition is that change points are correlated with a peak in community activity which would be detected as an increase in the dissimilarity gap between successive snapshots of the network. The gap to the (near) optimal flow would be a proxy of intensive change.

Other extensions to our work include the usage of Syntgen to evaluate community detection algorithms on temporal networks and analysing syntgen capabilities to reproduce empirical systems.

Chapter 4

Community identity in a temporal network: A taxonomy proposal

The following section was previously published in (Pereira, Lopes, & Louçã, 2021). There we proposed a classification of lifetime events communities can undergo. 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 that article we proposed 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”.

Synthetic networks generated by Syntgen as well as empiric networks were used to test and validate the classification and associated methods. Syntgen code functionality was further expanded with a back-end that analyzed and returned the communities evolution resulting from networks generated by the user seed specifications.

Finally, the community identity method was applied to a match of the game of soccer, where communities were defined as sets of interacting players. Their evolution was tracked and a statistical analysis of their lifecycle events was carried out. This initiated the transition from the more theoretically grounded topics of this thesis to real life applications over complex systems.

4.1 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 (Gordon-Roth, 2019).

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" (Loemker, 1969). 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 thesis 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 chapter, 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 (Fortunato, 2010). 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 (Barabási, 2015) 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 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 (Jaccard, 1912) 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 chapter we refer and review related work that predates our current proposal in section 4.2. In section 4.3 we describe how we compare communities to determine lifecycle events and their taxonomy tree. In section 4.4 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 4.5 we present the full classification methodology and procedure, followed in section 4.6 by examples using a toy model and an empiric network. We conclude in section 4.7 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.

4.2 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 (Jdidia, Robardet, & Fleury, 2009), 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(Aynaud & Guillaume, 2011), usually by establishing additional objectives, such as minimizing the clustering changes across time thru measures such as the normalized mutual information (Danon, Díaz-Guilera, Duch, & Arenas, 2005), 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 (Aynaud, Fleury, Guillaume, & Wang, 2013) 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 (Rossetti & Cazabet, 2018) authors expand on this classification, creating a hierarchy of approaches, that at the first level is similar to the one in (Aynaud et al., 2013), 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 thesis. The reader is referred to (Aynaud, Guillaume, Wang, & Fleury, n.d.; Dakiche, Benbouzid-Si Tayeb, Slimani, & Benatchba, 2019; Enugala, Rajamani, Ali, & Kurapati, 2015; Hartmann, Kappes, & Wagner, 2016; Rossetti & Cazabet, 2018; Spiliopoulou, 2011; Xie, Kelley, & Szymanski, 2013) for more information.

Although most of these efforts concentrate on identifying temporal communities in an absolute sense, in this chapter 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 (Palla et al., 2007) 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 (Rossetti & Cazabet, 2018). A summary of these events with informal definitions can be found in (Cazabet & Rossetti, 2019) 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 (Jaccard, 1912):

$$J(c_i^t, c_j^{t+\epsilon}) = \frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_i^t \cup c_j^{t+\epsilon}|} \tag{4.1}$$

is used by authors in (Greene, 2010; Mall, Langone, & Suykens, 2015; Nguyen, Kirley, & García-Flores, 2012; Palla et al., 2007), even though it may be named differently in some cases.

In (Takaffoli, Sangi, Fagnan, & Zaiane, 2010) 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 (Takaffoli, Sangi, Fagnan, & Zaïane, 2011) the same authors distinguish between communities and metacommunities, the latter being a construct to track community evolution. In (Asur, Parthasarathy, & Ucar, 2009), 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 (Hopcroft, Khan, Kulis, & Selman, 2004) authors use a measure that favors communities similarly sized with a high ratio of common nodes:

$$similarity(c_i^t, c_j^{t+\epsilon}) = min\left(\frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_i^t|}, \frac{|c_i^t \cap c_j^{t+\epsilon}|}{|c_j^{t+\epsilon}|}\right) \tag{4.2}$$

A different approach is taken in (Bródka, Saganowski, & Kazienko, 2013) 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 (Sun, Tang, Pan, & Li, 2015), 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 (Langone, Mall, & Suykens, 2015), 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 (Mall et al., 2015) (which we analyse further in section 4.6.2), 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 (Greene, 2010), 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.

4.3 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 cutoff. 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.

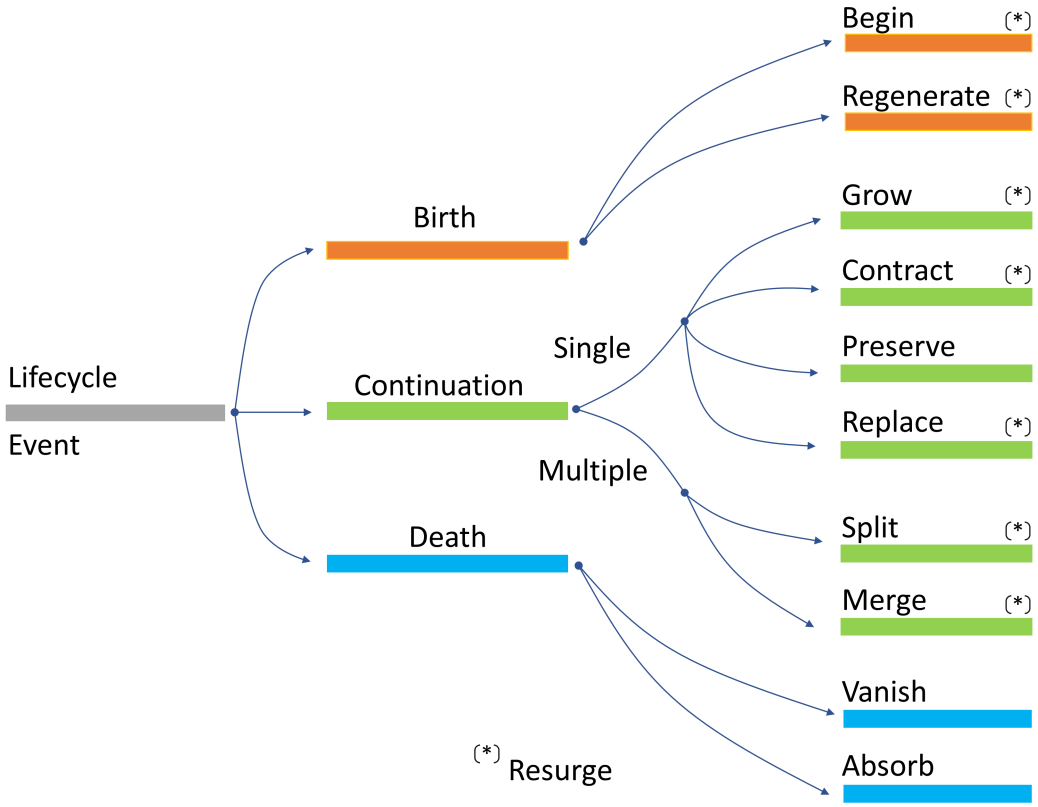


Figure 4.1. Events in the lifecycle of a community in a temporal network. Classification dependent on multiplicity of continuation events and relative set sizes

- 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 4.1. 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) (Jaccard, 1912), as previously mentioned. Authors in (Palla et al., 2007), 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.

4.4 Adjusted Jaccard index and null model

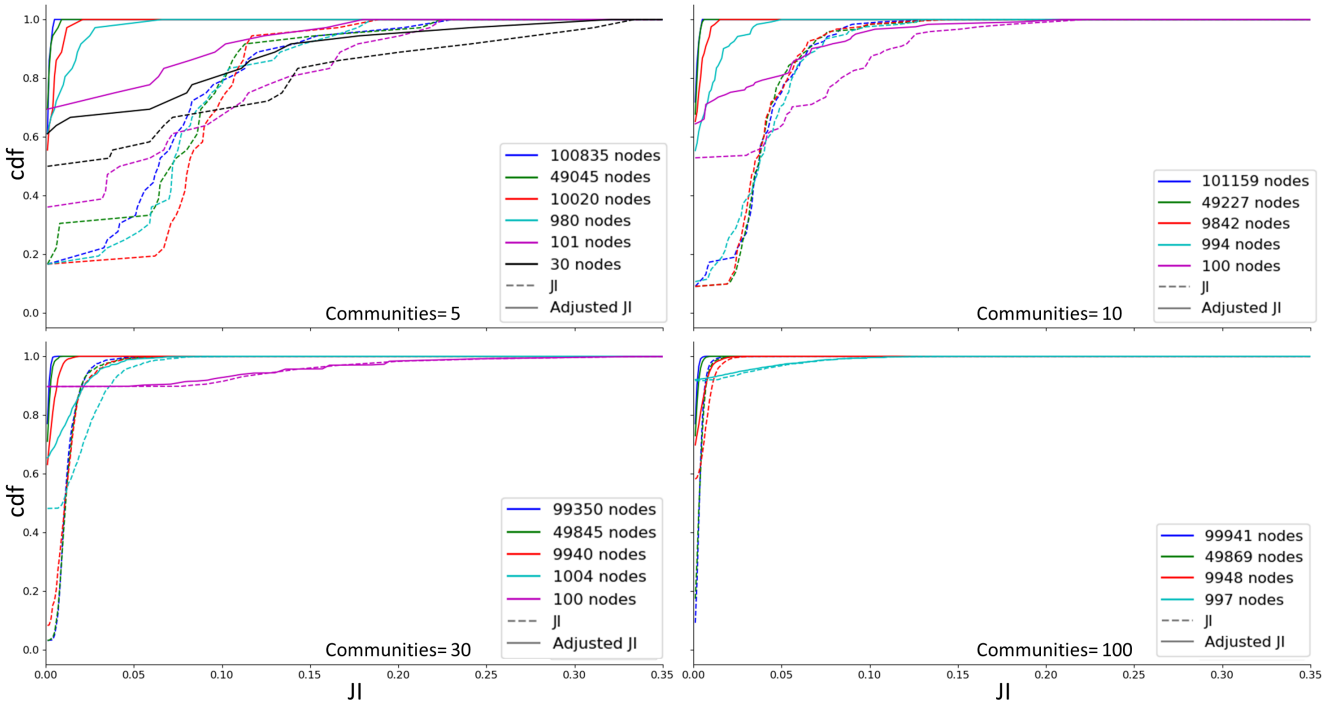


Figure 4.2. Performance of the Adjusted Jaccard index ($\hat{J}$) for the null model. Cumulative distribution function of $\hat{J}$ and J. Average $\hat{J}$ compared to averaged J for pairs of communities with a positive index for varying numbers of communities and nodes. Each line represents the average of 100 runs. As the number of communities increase, $\hat{J} \to J$.

A random network should not exhibit community structure[1]. Similarly, a random redistribution of community membership across the node set over $t \to t + \epsilon$, 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 $\breve{}$ accent.

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

$$\mathbb{P}(v \in \breve{c}_i^{t+\epsilon}) = \frac{s_i^{t+\epsilon}}{\sum S^{t+\epsilon}} \tag{4.3}$$

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

$$\mathbb{E}(|c_i^t \cap \breve{c}_j^{t+\epsilon}|) = s_i^t \times \frac{s_j^{t+\epsilon}}{\sum S^{t+\epsilon}} \tag{4.4}$$

[1]This fact is the basis of one of the most popular methods of community detection (Girvan & Newman, 2002)

[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.

for any community c^t, and the corresponding community $\breve{c}^{t+\epsilon}$ built from the probabilistic distribution of nodes onto $\breve{C}^{t+\epsilon}$ resulting from equation 4.3. Let's notate this $f_\emptyset(c^t, \breve{c}^{t+\epsilon})$, 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. $|c_i^t \cap c_j^{t+\epsilon}| \leq f_\emptyset(c_i^t, \breve{c}_j^{t+\epsilon}) \Leftrightarrow \hat{J} = 0$
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 $c_i^t, \breve{c}_j^{t+\epsilon}$, under the conditions of equation 4.4, basically the ratios of the intersection with the union of communities c_i^t and $\breve{c}_j^{t+\epsilon}$. We denote this index as $\breve{J}$:

$$\breve{J}(c_i^t, \breve{c}_j^{t+\epsilon}) = \frac{s_j^{t+\epsilon} \times s_i^t}{\sum S^{t+\epsilon} \times (s_j^{t+\epsilon} + s_i^t) - s_j^{t+\epsilon} \times s_i^t} \tag{4.5}$$

Formula 4.6 allows us to correct the index to zero on random chance, while preserving a perfect score of "1" when $c_i^t = c_j^{t+\epsilon}$:

$$\max\left(\frac{J(c_i^t, c_j^{t+\epsilon}) - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon})}{1 - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon})}, 0\right) \tag{4.6}$$

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:

$$r_{ij} = 1 - \frac{|c_i^t \cap c_j^{t+\epsilon}| - f_\emptyset(c_i^t, \breve{c}_j^{t+\epsilon})}{\min(s_i^t, s_j^{t+\epsilon}) - f_\emptyset(c_i^t, \breve{c}_j^{t+\epsilon})} \tag{4.7}$$

$r_{ij} = 0$ 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:

$$\hat{J}(c_i^t, c_j^{t+\epsilon}) = \frac{J(c_i^t, c_j^{t+\epsilon}) - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon}) \times R(c_i^t, c_j^{t+\epsilon})}{1 - \breve{J}(c_i^t, \breve{c}_j^{t+\epsilon}) \times R(c_i^t, c_j^{t+\epsilon})} \tag{4.8}$$

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

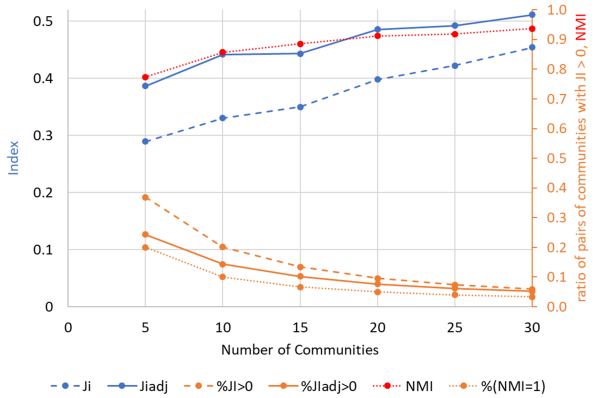
samples of such a network, ideally, it should result in *cdf*$(index) = 1 | index \in [0, \infty]$, with all observations at 0.

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 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 4.2. $\hat{J}$ 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 $\approx$ 100 communities and $\approx$ 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.

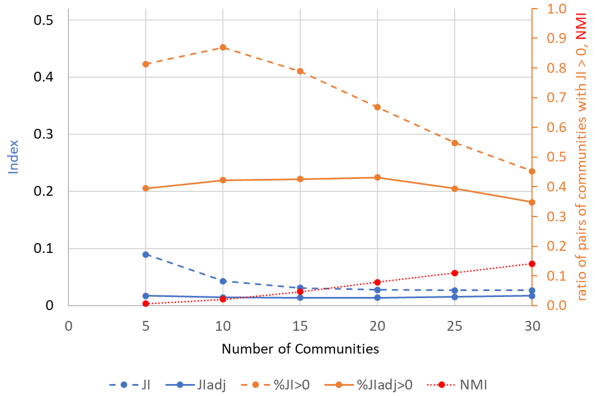
The larger divergence from the ideal behaviour on small networks is the result of two factors. With less nodes and less communities, the probability of spurious similarities increases as nodes have less degrees of freedom. That is taken care by the null model. However, discretization also plays a role 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 4.3a). These networks were generated using a tool (Pereira et al., 2020) 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 4.3 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 $\approx$ 1000 nodes, varying from 5 to 30 communities in steps of 5. In figure 4.3a 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 4.3b.

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 $\hat{J}$.



(a) Networks with low community volatility



(b) Networks with random communities

Figure 4.3. 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

.

4.5 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:

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 4.1.

3. A null Jaccard index matrix ($\breve{J}$) is created from $S^t, S^{t+\epsilon}$ using equation 4.5 .

4. An adjusted Jaccard index matrix $\hat{J}$ is created from J and $\breve{J}$ using equation 4.8.

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.

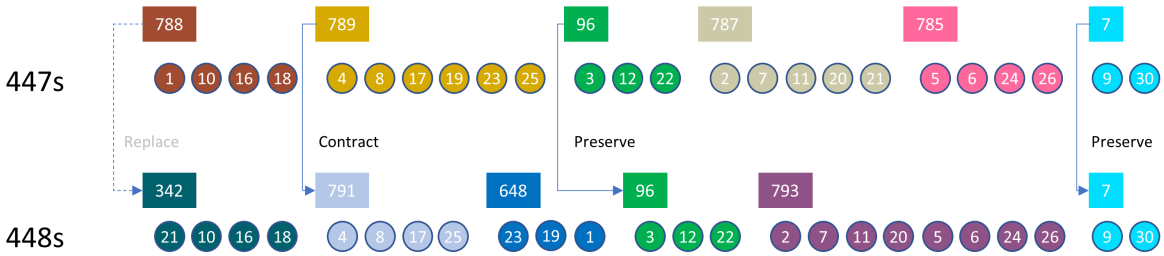
8. For every $h_{ij} = 1$ there is a multiple **continuation** top level event between communities c_i^t and $c_j^{t+\epsilon}$ if $(m_i \vee p_j) > 1$. Top level multiple **continuation** events generate second level:

   (a) **merge** events if $m_i > 1$;

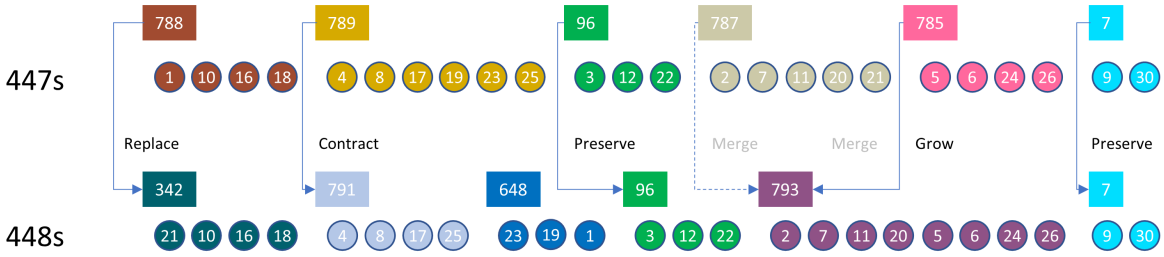
   (b) **split** events if $p_j > 1$.

   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) Community events at $\theta = 0.6$



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

Figure 4.4. 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.

(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.

4.6 Examples

In this section we present two examples of the application of the proposed taxonomy and method. In subsection 4.6.1, we use a toy model to illustrate the individual steps taken to determine community lifecycle events, and, in subsection 4.6.2, 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.

4.6.1 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 section 4.5):

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

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):

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

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

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 **absortion** 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.

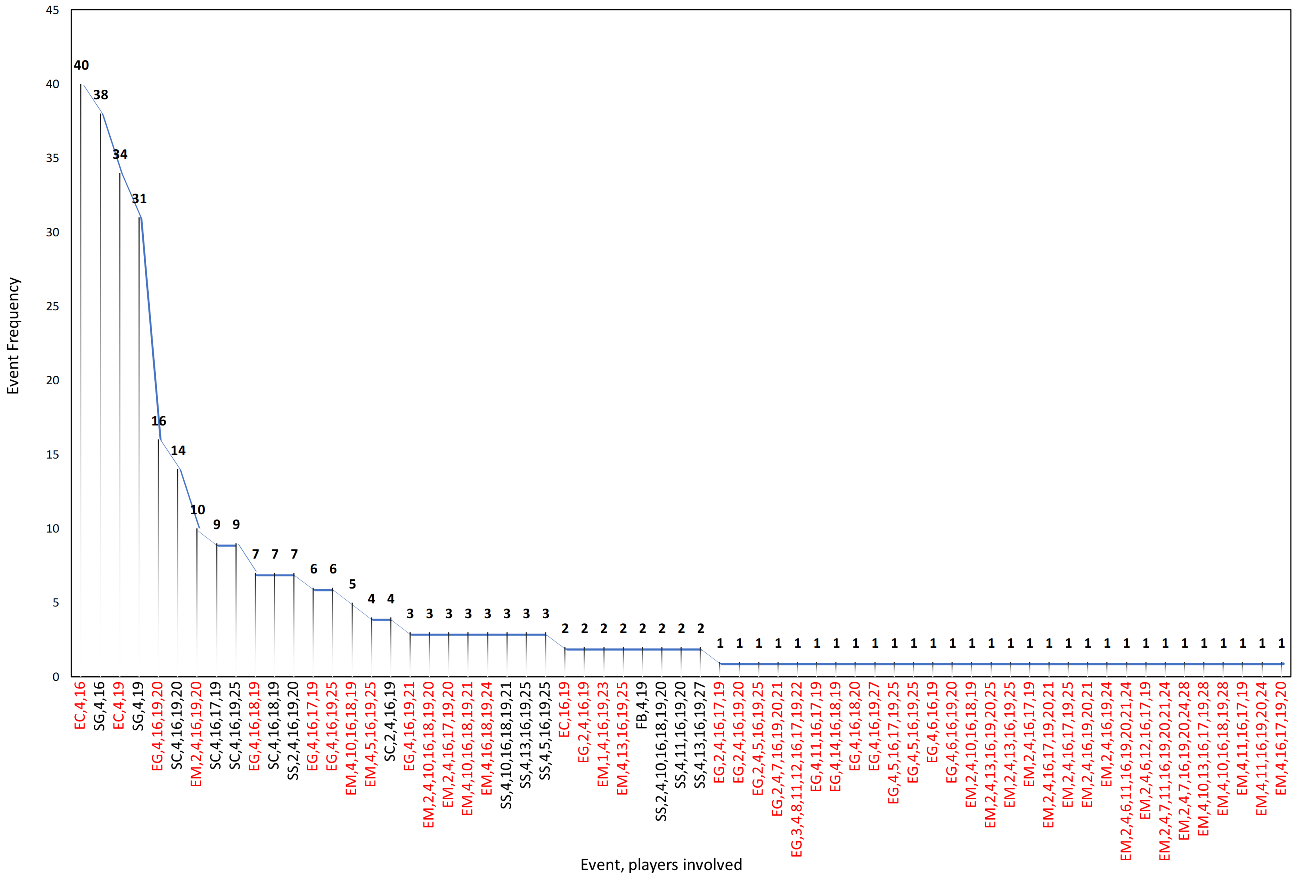
4.6.2 Application to an empiric network

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 (J. Ramos, Lopes, Marques, & Araújo, 2017). The match is sampled at 10Hz, generating close to 60,000 observations during the whole game. 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 (J. P. Ramos, Lopes, & Araújo, 2020). 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 (J. Ribeiro et al., 2019). 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 4.4. Seen thru the "lens" of our method we can see the events the sets of players underwent. At a cutoff $\theta = 0.6$ (figure 4.4a), 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



Type	Key	Event	Comments	Type	Key	Event	Comments
Exit	EA	Absorption	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 4.5. 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.

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 4.4b we relax the cutoff point lowering it to a level ($\theta = 0.42$) where clear differences to figure 4.4a 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 (Mall et al., 2015) 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 4.5). 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 (J. Ramos et al., 2017). Set 63 (players 4, 16, 19) is such a set in the match data we are using for this example. In figure 4.5 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.

4.7 Conclusion

In this chapter 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. 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.

4.8 Supplementary Material

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

Chapter 5

The Soccer Game, bit by bit: An information-theoretic analysis

The following section was previously published in (Pereira, Lopes, Louçã, Araújo, & Ramos, 2021). As a last step in the journey presented in this thesis, the contributions described previously were used to study the soccer game as a complex network system that exhibits dynamic clustering behavior. We had access to data from a major European national football league, that comprised players' coordinates sampled at 10Hz. We measured the rate of positional change based on clusters that were assembled from the relative spatial distribution of players on the pitch. The rate itself was quantified applying the method used in Syntgen to compute differences between network partitions, understood as the set of communities observed on a sample. We found significant correlations between measurements and key match events that are empirically known to result in players jostling for position, such as when striving to get unmarked or to mark. These events increase the amount of change between samples, while breaks in game play have the opposite effect. Having a measurement of dynamic, structural change in soccer is an original contribution that can complement full match statistical analysis. Another significant benefit of this approach is its ability to hierarchically decompose measurements at multiple levels, building an overall multi-layer map that provides insights into the game dynamics, from the individual player, to the clusters of interacting players, up to the teams and their matches. This comprehensive view of the players' interacting behavior can be useful for training, tactics and strategy development.

5.1 Introduction

Complex systems, with time evolving interactions among its elements, abound in the social, biological and physical domains. In many of these systems, elements are clustered in groups that also undergo changes with time. A temporal, clustered network can be an appropriate representation of such a system.

In this chapter we apply this representation to the sport of soccer. Soccer, as many other competitive team sports, can be seen as a socio-biological complex system. The domain dynamics of agent behavior in these sport modalities are neither fully random nor fully designed (J. P. Ramos et al., 2020). This contributes decisively to their complexity. Agents cooperate and compete in clusters towards shared and conflicting goals. These clusters are frequently functionally bounded, such as in the group interactions of forward and defense players, or goal keepers and strikers. It is common knowledge that self-organization in complex systems emerges from constrained local action, so this representation appears, in principle, justified. A comprehensive discussion of the application of complex systems theory to football can be found in (Salmon & McLean, 2020).

The soccer match is represented in this chapter as a succession of network observations where clusters are subsets of players, including the two football goal frames, resulting in a network with a maximum of 24 nodes concurrently active, plus substitutes (J. Ramos et al., 2017).

While studying a soccer match as an evolving clustered network, we start from the proposition that players' spatial distribution is the determining variable for clustering. Not in relation to the pitch boundary, but in relation to their teammates and adversaries. This is reasonable if we consider that being marked or unmarked, supported or unsupported, has a major impact on the opportunities for action that a player enjoys. The research question that this study aims to answer is whether the changes these clusters suffer as the game evolves, promise a more faithful indication of game dynamics when compared to traditional measurements such as the ratio of successful passes, speed, distance covered, and others.

Intuitively, we could think that an optimal assignment of players to clusters would require a physical distance measure, predicating link weights by player relative distance. However, there are complicating factors to the usage of such a precise measurement, as the importance of inter-player distance is not independent of game play (J. Ramos, Lopes, & Araújo, 2018). It varies with pitch location, ball position, game rules, environmentals (such as playing surfaces or weather), or the relation between time and distance in dynamic game settings. All these contribute to the actual player's instantaneous grasp of his performance environment and perception of opportunity for action (Araújo & Davids, 2016).

This was the basis that lead us to cluster players and goals into homogeneous and disjoint groups connected by a single link (J. Ramos et al., 2017), using the formalism of hypergraphs (Berge, 1973). A hypergraph is characterized by having multiple nodes connected by a single link, called a hyperedge, in contrast with a traditional graph where links have a maximum of two endpoints. A set of nodes that share a link is called a simplex. In our particular context, simplices are sets or clusters, and the collection of simplices observed in a single sample, a clustering.

We use the term "clustering" to mean the set of disjoint non-empty subsets of nodes observed in the network at a given point in time. Some authors call it a "partition". These terms

represent similar constructs, clustering being semantically associated with an emerging, bottom-up aggregation of nodes, while partition conveys the idea of a top-down driven process. In soccer there is not a single entity controlling group formation (J. Ribeiro et al., 2019), at least not directly and in real time, so the former seems more appropriate.

In the restricted context of this chapter, simplices and clusters are synonyms, both referring to the same construct: a group of players in articulated interaction and proximity. An example of the clustering process is illustrated in figure 5.7, and the actual method steps in algorithm 3. Both can be found in the appendix of this chapter.

It could be argued that discretization and assignment of nodes to a pairwise disjoint family of sets, would lead to a distorted representation of events on the pitch. After all, players move freely in an Euclidean space and in continuous real time, while in the proposed representation time is discrete and players move on a lattice, understood not as a grid that spans the pitch but as the configuration space of all possible set arrangements (Conway & Sloane, 1999; J. Johnson, 2010). Frequent observation, however, mitigates these effects. For example, peripheral players in a simplex will more easily transfer to a different simplex and, if frequently observed, any simplex changes will be quickly captured. Due to the high frequency characteristic of the network (10Hz), errors will smooth out as player simplices form and dissolve, establishing a bridge between the continuous domain of game play and the time sliced network representation employed (J. H. Johnson, 2016).

This discretization carries with it a significant advantage. We are no longer in a continuous domain, and the toolkit of information theory (T. J. Cover & Thomas, 2006) becomes available to us. In a discrete domain, information can be quantified for complexity, such as in the Kolmogorov complexity or the Shannon entropy (Grünwald & Vitányi, 2008; Kolmogorov, 1968; Shannon, 1948).

Similarly, two pieces of information can be compared for distance. We can determine how far apart or how close they are by the number of units of information that are needed to find one given the other. The pieces of information are the individual clustering samples of the soccer match. We measure their distance using the *Variation of Information*, an entropic based metric introduced by Marina Meilă in 2003 (Meilă, 2003), to compare clusterings. A detailed description and reasons for selection can be found in section 5.3.1. It's on this intersection of network science and information entropy that this chapter is rooted.

In the reminder of this document, we discuss related work in section 5.2. Theoretic underpinnings, including major theories, concepts, key variables and the way they inform observations, correlation of VI and playing dynamics and procedures used are in section 5.3, which is followed by a section 5.4 describing our findings. We discuss these results in section 5.5 and we conclude with directions for future research in section 5.6.

5.2 Related work

Using networks and entropic measures to study the soccer game is not new. In this section we refer to prior studies that have explored these techniques and explain how they differ from this chapter's approach. This is not a comprehensive review or description of networks or entropy and their use. The reader is referred to (J. Ribeiro, Silva, Duarte, Davids, & Garganta, 2017) for a summary of the implications and merits of applying network science to team sports performance analysis, and to (M. Ribeiro et al., 2021), where a description of the extensive variants of entropy, some of which have been used in team sports analysis, can be found.

In comprehensive reviews of the literature, such as those found in (Lord, Pyne, Welvaert, & Mara, 2020; Sarmento et al., 2018) where authors analyze performance and general research trends in soccer and other team invasion sports, networks are a popular topic. In (Sarmento et al., 2018), a review fully dedicated to soccer, 11.7% of articles reviewed use networks and network metrics as an analysis tool, and in (Lord et al., 2020), a review of the literature on performance analysis of team invasion sports, 10.8% of the reviewed articles focusing on soccer make use of network analysis. All of these articles use exclusively networks built out of dyadic interactions between players on ball passing and crossing, sometimes incorporating spatio-temporal metrics (Clemente, Martins, & Mendes, 2016; Cotta, Mora, Merelo, & Merelo-Molina, 2013; Gama et al., 2014). Usually a weighted digraph is built per team, sometimes broken down to individual attacking play (Korte, Lames, Link, & Groll, 2019), and statistics such as clustering coefficient, network density, centrality or degree distribution are used to explain patterns of play or performance. Spatial analysis is accomplished dividing the pitch into diverse zones, either longitudinally or on both axis, and assigning arcs (i.e. directed links) connecting the passes' origin and target zones. Specific attacking plays, such as those ending up in a scored goal have also been analyzed using these techniques (Mclean, Salmon, Gorman, Stevens, & Solomon, 2018).

Entropy has been previously used to study soccer dynamics, but much less frequently than network science. As an example, in the reviews referenced above, there are only two explicit references to articles dealing with soccer and entropy.

In (Vilar, Araújo, Davids, & Bar-Yam, 2013) authors clustered players by their location in seven pitch sectors, dynamically bounded by the 20 outfield players. Similarly to our approach, this clustering is performed every 0.1s. They then computed the difference in the number of players from each team in each of the sectors and measured the Shannon entropy of its frequency for the whole match, resulting in an uncertainty measurement of local dominance. This was used to identify correlates of performance and patterns of intra and inter team coordination, understood as the level of sector numerical dominance that results from player interactions. Although the sectors, and thus the clusters, are dynamically defined, there is a level of inflexibility by fixing the number of clusters of players per observation. The clustering method also does not avoid assigning players in closer interaction to separate clusters. In contrast with the entropic

measure used in this thesis, it prevents fine grained temporal analysis, as it is frequency based.

In (Lopes & Machado, 2019) Shannon entropy (among other information theoretic measures) is used to study multiple national leagues using rounds as time units, with home and away goals as variables. The authors found the emergence of similar entropy patterns across seasons and across leagues.

In (Couceiro, Clemente, Martins, & Tenreiro Machado, 2014) authors quantified space coverage variability of players, by discretizing the pitch area into 1 m^2 cells and using the frequency distribution of players over the cell map to compute its Shannon entropy. As expected they found that midfielders exhibit a higher entropy than other players. According to the authors, this result is more "assertive" than a typical heat map. Approximate entropy, a time series analysis technique that can reveal the predictability of patterns, was also used in this chapter, to analyze the distance covered by a defender. It was possible to categorize the respective time series (at 1s interval) as a chaotic system, somewhere in between periodic and random.

Approximate entropy, was also used to analyze spatial statistics, such as occupied areas, dispersion or team center of gravity in (Duarte et al., 2013). No clustering of players was performed and time analysis was limited to 15 min segments. The same technique was also used in (Sampaio & MaçÃs, 2012), with different spatial and dynamical properties, to study the effect of tactical training in a group of student footballers playing small sided matches.

In (Martínez et al., 2020; Y. Neuman, Israeli, Vilenchik, & Cohen, 2018; J. Ribeiro et al., 2020) we find examples of studies that use networks in conjunction with entropic measures in match analysis. In (J. Ribeiro et al., 2020), authors used the same network formalism and representation as used herein, and sample entropy to measure the synchronization between players, their simplices and teams, from a time series of observed cluster phases. They observed different axial synchronization of player-simplex phases, on two small sided games setups with different conditions of goals', number, sizes and location (4 mini-goals without goalkeepers versus 2 larger goals with goalkeepers).

(Martínez et al., 2020; Y. Neuman et al., 2018) are both based on pass networks. In (Y. Neuman et al., 2018) authors used the Tsallis entropy, a measure that generalizes the traditional Boltzmann-Gibbs/Shannon entropy to non-extensive systems (that is, systems where sub-states are not mutually independent), to study its correlation with team performance and season results. The analysis is performed at match level, and the authors found that, under certain parametrization, the Tsallis entropy of a team is inversely correlated with team performance. The opposite result is observed when considering the difference of team entropies per match. In (Martínez et al., 2020) authors performed a spatial entropy analysis of pass origins at match level, and a temporal analysis of network parameters with high correlation with the number of passes, such as the longitudinal coordinate of center of mass of the pass network or the network clustering coefficient, using permutation entropy on a time varying series built with a moving window of 50 passes.

The network design approach we took for this chapter diverges substantially from a passing

network. It is self-evident that only a player in possession of the ball can score, which is a strong argument in favor of using passing networks for performance analysis. However, as pointed out in (Grund, 2012), relevant interactions in a soccer game are not limited to passes. Intuitively, the opportunity for a successful pass is perceived by the player carrying the ball, as a function of multiple variables, in which the dynamic position of some of his teammates and adversaries play a major role. The same can be said for the opposing team while trying to intercept or clear a pass. As mentioned in (Hewitt, Greenham, & Norton, 2016) "Players must be able to pass with precision while others create space around themselves to receive the pass from their teammate". It is dynamics like this that we try to capture by using the formalism previously introduced. In the specific case of passes, the temporal changes in clusterings are precursors for a passing opportunity or interception. In non formal language, we can say that in a passing network we can find what happens, while in a polyadic network of player's interactions, we can explain why it happens!

There are other differences in the proposed approach that circumvent some of the challenges of passing networks. Relations in passing networks are inherently dyadic, although, as mentioned, a player passing decisions are inherently polyadic. Passing networks are usually a single team view, where the influence of the opposing team is usually absent. Interceptions and clearances are ignored, although they may have a decisive impact on the game. The use of signed networks could address some of these difficulties, but introduce theoretical challenges, as many of the metrics of simple networks do not extend to signed networks, which is probably the reason that, to our knowledge, they have not been used for this purpose. And, finally, compared to positioning actions, passes are relatively rare, leading to a low temporal resolution when gathering statistics. In (Yamamoto & Yokoyama, 2011), authors propose a minimum window size of 5 min, to collect passing data. The reader is referred to (Buldú et al., 2018) for a thorough discussion of the challenges of using passing networks.

The representational formalism used in this chapter was introduced in (J. Ramos et al., 2018, 2017). In those articles, every match observation was partitioned into clusters of players in proximal spatial interaction, and several variables were extracted from this representation. Here, we extend this prior work to reveal the changes these clusters experience across time, and explore their meaning by using an information entropic metric.

In summary, the major original contributions introduced in this chapter and detailed ahead, are:

- Using dynamic polyadic relations between players, more faithfully representing the player decision making process
- Measurements of cluster breakup and emergence that encompass home and away teams and their dynamics
- Structural change measurements that can be decomposed at multiple levels

- Change measurements without a fixed frame of reference, avoiding some of the pitfalls of traditional measurements.

5.3 Methods

In this section we cover the theories, concepts, constructs, key variables, and the way they inform the observations in section 5.3.1, and the procedures used to represent and analyze the captured data from the sample set of matches in section 5.3.2.

5.3.1 Theoretical Framework and Underpinnings

Every observation of a match is a clustering of nodes, representing players and goals. Formally, a clustering is:

$$C = \{c_1, \cdots, c_k\} : (c_i \cap c_j = \varnothing \quad \forall \ (1 \leq i, j \leq k \ \ \wedge \ \ i \neq j)) \wedge \ \cup_{i=1}^{k} c_i = V \quad (5.1)$$

where c are the disjoint subsets, k the number of subsets, and V the set of all nodes.

There are several methods to measure the inter-distance between clusterings, with varying properties, such as the Rand Index (Rand, 1971), Adjusted Rand Index (Hubert & Arabie, 1985), the Normalized Mutual Information (Danon et al., 2005), the Van Dongen-Measure (Dongen, 2000) and others. A thorough discussion of the major methods can be found in (Meilă, 2007; Vinh, Epps, & Bailey, 2010; Wagner & Wagner, 2007). We selected the Variation of Information (VI) (Meilă, 2007), also known as Shared Information Distance, to measure the information distance between samples and thus evaluate the change a clustered network experiences as a function of time. The choice of VI is justified as it is a true metric, respecting the triangle inequality, meaning that no indirect path is shorter than a direct one. This is important in analyzing the rate of change at multiple scales, avoiding the unreasonable possibility of having a greater rate of change for a given time interval, when sampling the network at a lower rate. VI also increases when fragmentation and merges occur in larger clusters, which intuitively relates to playing dynamics, given the rise in degrees of freedom experienced in larger groups of interacting players. Fundamentally, although in this chapter we consider VI as a proxy for game dynamics, VI itself is not a quantification of informational meaning or semantics, but simply, a quantification of informational variation, or as Shannon puts it "semantic aspects of communication are irrelevant to the engineering problem" (Shannon, 1948).

In simple terms, VI, measures the amount of information required to obtain one clustering (observation) from another. If no changes in the clusters are observed, then there is no variation of information. As clusterings shift from one another, VI increases. This is easy to visualize when considering the so-called confusion matrix (Stehman, 1997) between clusterings at successive observations. This matrix describes the node spread, where each element represents the

number of nodes moving from one cluster to another. If clusters are unchanged and keep their node affiliation, the confusion matrix will be a monomial matrix, $VI = 0$ and we know exactly where each node ends up. But as the number of non-zero entries in the confusion matrix increases and their distribution tends to uniform, the uncertainty about each node destination also increases. Consider as an example a cluster that splits in half versus another that sheds a single node. There is a higher uncertainty about each node final destination in the former than in the latter. VI measures this uncertainty. A practical illustration of how to compute VI can be seen in tables 5.2 and 5.3 in the appendix.

Formally, VI is a function (see equation 3.6) that takes two clusterings as parameters and returns the information distance between the clusterings. From this equation it is easy to see that when the clusters in X and Y are the same, the result is zero, as $r_{ij} = p_i = q_j$. This result expresses the fact that no information is gained or lost when going from one clustering to the other. For empty intersections of pairwise clusters, $r_{ij} = 0$, and although $\log(0)$ is not defined, applying l'Hopital rule we get a null contribution from these intersections to the overall VI. In summary, only pairwise non-disjoint, non-identical clusters contribute to the information distance. This contribution led us to introduce an additional construct, the simplex transition. Simplex transitions can be statistically analyzed, and their frequency and contribution to overall VI, can provide insights into structural change and dynamics of the match.

VI works as a distance metric for clusterings of the same set of nodes. In the model used to represent the soccer match, the set of nodes remains constant, except on substitutions and send-offs. However, the number of observations affected by these events are so low, that we have ignored their contribution in the model.

Using base 2 logarithms, VI is measured in bits (or shannons) and describes the balance of information needed to determine one clustering from another. VI is algorithmically simple (it can be computed in $\mathcal{O}(n + kl)$)) and, as mentioned before, it is a true metric (Kraskov et al., 2005a), respecting positivity, symmetry, and the triangle inequality.

Using the previous notation, for every individual player $p_{ij} \in \{x_i \cap y_j\}$ his contribution to the overall VI is computed as:

$$VI^{p_{ij}} = -r_{ij}\frac{[\log_2(\frac{r_{ij}}{p_i}) + \log_2(\frac{r_{ij}}{q_j})]}{|x_i \cap y_j|} \tag{5.2}$$

which takes the contribution of pairwise clusters x_i, y_j to the overall VI, and divides it in equal parts among all players $\in x_i \cap y_j$. Note that, in the particular case of the network that we built, all nodes/players are present in all observations and are members of one and only one cluster in any one observation. Equation 5.2 registers the contributions of players involved in their clusters when these change. The only exception is the case of a send-off or substitution, in which case the player no longer contributes to the dynamics of the match.

The VI of two clusterings (X, Y) of S can only be zero if $\forall s \in S \mid s \in X \leftrightarrow s \in Y$. If this condition is not met then $\min(VI) \geq \frac{2}{n^*}$ (Meilă, 2007), where $n^* = \max(k, l)$ still using the

same notation. In the soccer match representation here proposed the number of nodes is fixed at 24 (barring any red cards), and thus, $n^* = 12$ and $\min(VI) = \frac{1}{12}$ every time there are any clustering changes. VI depends on the level of fragmentation on the pitch across observations, which intuitively reflects the situation of players jostling for position, but cannot exceed $\log_2(n)$ (Meilă, 2007). These extreme values of VI are, however, just boundaries that limit minima and maxima given any set of clusterings. In the present case, we have a minimum of 2 nodes per cluster, which implies a maximum of 12 clusters, resulting in $\max(VI) = \log_2(12) = 3.585$, which is attained when a clustering with a single cluster splits into 12 clusters with two nodes each, or vice-versa. In practice, the maximum VI registered is substantially lower with typical observed values of $\max VI \approx 1.2$, corresponding to the maximum distance between clusterings with $0.1s$ separation.

5.3.2 Procedures

The proposed framework was applied to the analysis of a set of 9 soccer matches from the 2010-11 season of the English Premier League. Based on an information stream collected from realtime pitch-located raw video feed, each match is modeled as a high-resolution (10Hz) temporal hypernetwork with simplices as clusters of players and goals parsed by proximity. Each network is made up of up to 30 nodes (28 players and 2 football goals) of which only a maximum of 24 are present on the pitch at any given moment (11 players from each team and 2 goals). The inclusion of goals is justified when considering that the purpose of the polyadic formalism that we use is to capture the multiple factors that may affect a player's decision making process, and proximity to goals is certainly an important one. The number of simplices is variable, dependent only on the observed map of players and goals. The method used for clustering guarantees that a node and its closest node belong to the same simplex, or, in other words, it guarantees that no node is closer to a node belonging to a different simplex than to its closest node in the same simplex. This implies that the smallest simplex has a minimum of 2 nodes, i.e., there are no isolated nodes. Although there maybe occasions where a player is side-lined, this will be an exception, as the expectation at the top-level of sports performance is that every single player have an active role in-play, in relation to their teammates and their opponents. Although the football goals are obviously fixed on the pitch, there is no fixed frame of reference for the clustering process. The algorithm used for clustering is non parametric and is explained in (J. Ramos et al., 2017).

On average, considering a match, including extra time, we observed and measured the network $\approx$ 60,000 times. Each of these 60,000 samples is a clustering of the network.

The output of the method is a time series of VI measurements, that can be hierarchically decomposed into separate measurements for teams, players, and simplex transitions.

At 10Hz, a significant amount of sparsity, i.e. a large amount of transitions without clustering changes, is observed. This posits the question of the ideal sampling rate (Moura et al., 2013),

given the dynamics of a soccer game, the capturing technology and the clustering methodology. The observed sparsity lead us to adopt a set of measures in the findings section ahead, to enhance analysis and observability. These included:

- the usage of differentials and measuring change in bps, denoted as $\dot{VI}$;
- the use of moving averages for visualization and compatibility with the rate of change and play of a soccer match. Results shown use 4s sample windows, except when noted;
- and, finally, we made use of cubic Hermite splines (E. Neuman, 1978) to envelope $\dot{VI}$ maxima. Results use an inter pivot distance that dynamically varies up to a maximum of 80s depending on the position of the observed value in the probability density function of $\dot{VI}$ (figure 5.1).

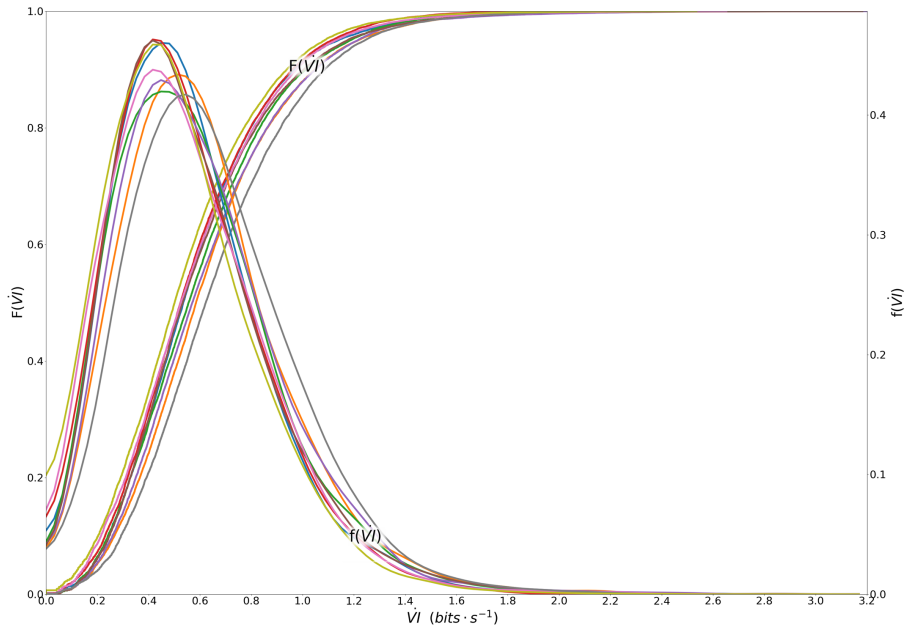


Figure 5.1. Probability Density Function ($f(\dot{VI})$) and Cumulative Distribution Function ($F(\dot{VI})$) for all nine matches measured on a 4s moving average window. Games color coded. There is a consistency of patterns that likely mirrors energy expenditure and management throughout the game (Osgnach, Poser, Bernardini, Rinaldo, & Di Prampero, 2010).

5.4 Findings

In this section we present the key findings resulting from our analysis.

5.4.1 Clusterings reappear much more frequently than expected by chance

Given that the space of all clusterings is substantial, corresponding to a lattice of over 4.4×10^{17} points (Bell number B_{24}), the amount of unique clusterings we can observe is just a small

fraction of this space, gated by the total of samples collected (average 58283, $\sigma = 1336$). Assuming a random distribution, the probability of observing the same clustering, that is the same sets of simplices, is for all purposes nil when considering the space size (1.46×10^{-12}). Obviously the real distribution is not random and is heavily condition by its prior state. But, when excluding consecutive observations, a significant level of clustering re-appearances still emerges (average 6.4%, $\sigma = 0.5\%$), which, intuitively, can be interpreted as the influence of strategic design over match playing patterns (J. P. Ramos et al., 2020).

5.4.2 Different time series, similar statistics

Having analyzed nine soccer matches of the 2010-11 season of the English premier league at 10Hz, on a 40 sample moving average window (4s), we found that the average $\dot{VI}$ and the standard deviation for the whole match is consistent across matches, with a total average of 0.597 bps, $\sigma = 0.0369$.

Considering that a typical player spends on average over half of his time standing or walking and only sprints ($> 8.3ms^{-1}$) 1.4% of the time (Ferro, Villacieros, Floría, & Graupera, 2014), 10Hz is a sampling frequency that often generates no clustering changes in consecutive samples. In fact, in almost 80% of the network observations clusterings do not change. The standard deviation per match has an average $\dot{VI}$ of 1.30 bps, with a maximum of 1.37 and a minimum of 1.25 bps across all nine matches. A full report for all matches can be found in table 5.4.

The dispersion of $\dot{VI}$ as measured by the coefficient of variation of all match observations averages 218%, reflection of the high activity level of the soccer game.

We found no correlation between the time ordered sets of VI observations between the matches we have analysed. When comparing different matches, we found consistent $\dot{VI}$ averages, with a coefficient of variation of the averages of $\approx 5\%$.

The probability density function of a match $\dot{VI}$ measurements is highly consistent across matches as seen in figure 5.1. Matches exhibit similar probabilities of finding given levels of dynamics and we did not find matches where $\dot{VI}$ is consistently high or consistently low. An explanation is player's regulation of exertion during the match to manage fatigue, particularly at the high intensity professional matches are played (Sarmento et al., 2018). All matches come from the official English premier league games, usually played at a similar competitive level, so these results are not surprising, if $\dot{VI}$ does accurately reflecting game dynamics.

5.4.3 Time decreasing trend of VI

In 8 out of the 9 matches we examined, we observed a lower VI when comparing the second half to the first half. Neuromuscular, biochemical and perceptual changes leading to increased physical and mental fatigue as a match progresses has been extensively documented (Silva et al., 2018). More specifically, indicators such as total distance with the ball, high intensity running with the ball, among other typical indicators of performance have been shown to measure

lower on the 2nd half of a match (Rampinini, Impellizzeri, Castagna, Coutts, & Wisløff, 2009). Adjusted tactics, resulting from increased acquaintance with competitor behavior, may be a further compounding cause.

A reduction in physical match performance (high speed running and sprinting) has also been reported when comparing the first 15 minutes of the first and second half (Weston et al., 2011). In line with this report, in our sample we observed a lower VI in all matches, under the same conditions.

However, it is important to note that in our sample the same team plays in every match. A larger sample of matches, from a wider population, may offer more consistency to this pattern, although these results already suggests a strong correlation between the proposed metric and game intensity, deserving further study. The observed values of VI can be seen in table 5.1.

Table 5.1. Comparison of $\dot{VI}$ between 1st and 2nd half of 9 matches, and between first 15 minutes of each half. In only one match do we observe values (shaded red) contrary to reported trends of intensity indicators

Match	1st Half	2nd Half	15 min 1st Half	15 min 2nd Half
1	0.555	0.533	0.566	0.533
2	0.611	0.571	0.601	0.532
3	0.634	0.628	0.703	0.623
4	0.679	0.650	0.703	0.693
5	0.614	0.630	0.630	0.624
6	0.590	0.556	0.698	0.584
7	0.599	0.539	0.617	0.518
8	0.639	0.559	0.639	0.546
9	0.603	0.558	0.550	0.547

5.4.4 Notational event data correlates with VI

To validate the hypothesis that VI is a measure of game dynamics, we searched for correlations between known moments of intensive player repositioning and surges in the information distance. Corners, being overwhelmingly defended one-to-one (Pulling, Robins, & Rixon, 2013), result in quick player displacement and occur frequently in a match (mean 10.3, $\sigma = 2.5$, which matches previously reported numbers (Casal, Maneiro, Ardá, Losada, & Rial, 2015)). This justifies, in our view, the selection of corners for hypothesis validation.

We collected timed tags for corners from match commentary. These events are time tagged down to the minute of play. To address the different resolutions scales of commentaries and clustering samples, we computed, per match, the mean $\dot{VI}$ for every minute of play, and compared its median with the mean for the minutes when corners were taken. Out of 93 corners, 86 had a higher VI than the median. The probability of this occurring on random chance is 1.33×10^{-16}.

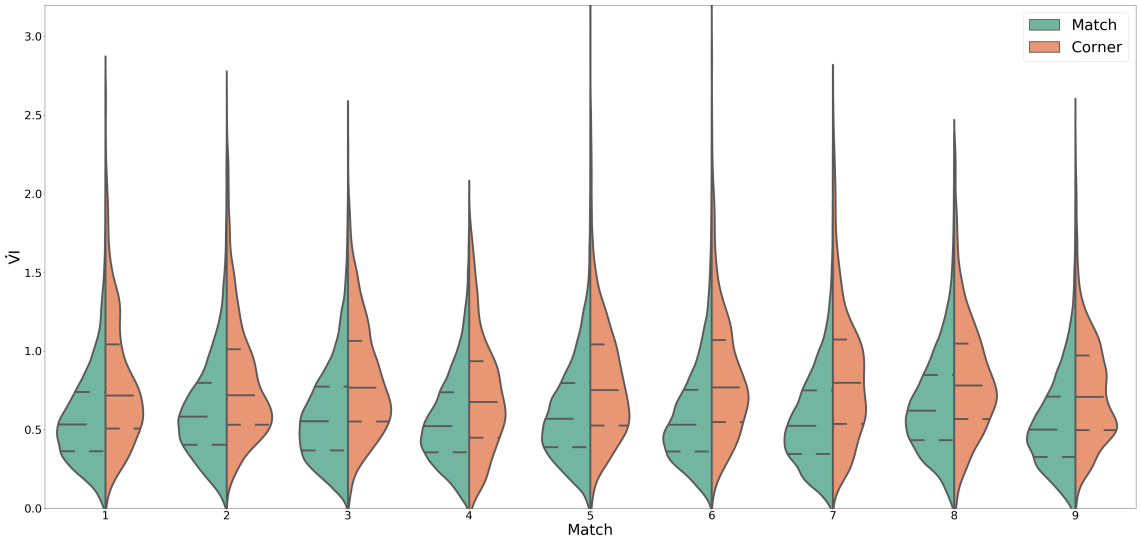


Figure 5.2. Distribution of $\dot{VI}$ for a complete match, compared with observations during the minutes when corners are taken, using a Gaussian kernel density estimate. There is a notable skew towards higher $\dot{VI}$ observations during corners for all nine matches.

We also inspected the $\dot{VI}$ distribution for the whole match and compared to the same distribution for all corners' minutes. As can be seen in figure 5.2, all matches show a $\dot{VI}$ distribution that is skewed higher when comparing with match averages.

These results provide compelling evidence that corners do indeed result in a marked increase of VI. VI, as used in this study, is clearly a proxy for game dynamics, understood as a rapid pace of inter-players relative displacement, i.e. without a fixed frame of reference. This is notably obvious during set pieces. Corners and free kicks invariably generate a spike in VI. Conversely, other events, like substitutions or send-offs, generate pauses that are captured by a drop in VI. Examples can be seen in figure 5.3a and 5.3b, where $\dot{VI}$ is plotted for a whole match, with vertical bars indicating the type and time of events.

5.4.5 Most simplex transitions occur only once

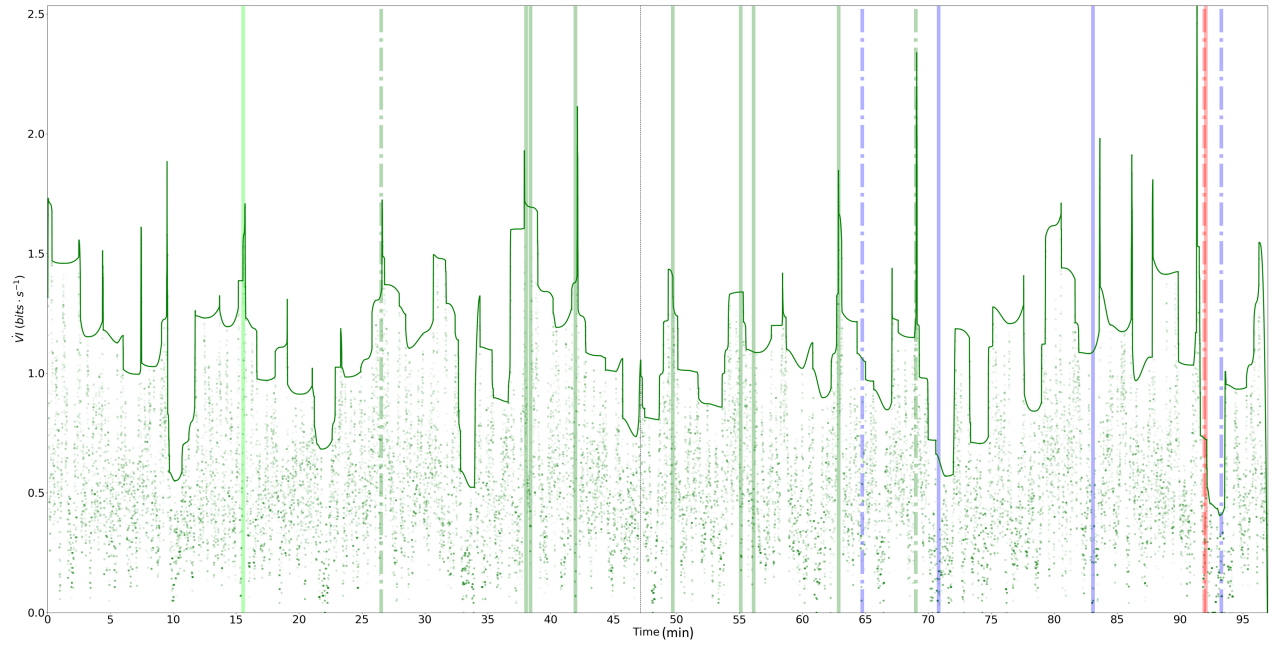
We also introduce the concept of a simplex transition, a tuple of simplices $(c_i^t, c_j^{t+\delta})$ such that $(c_i^t \cap c_j^{t+\delta} \neq \varnothing) \wedge (c_i^t \neq c_j^{t+\delta})$, that, at successive observations, involves always the same players.

Most simplex transitions occur only once during a match. However there are some that occur with higher frequency (up to 50 times a match). These are usually symmetrical. They may be candidates for further analysis given their relative importance. In figure 5.6 the top contributing transitions of one match are represented, indicating their relative $\dot{VI}$ weight, the nodes involved, and when during the match they occurred.

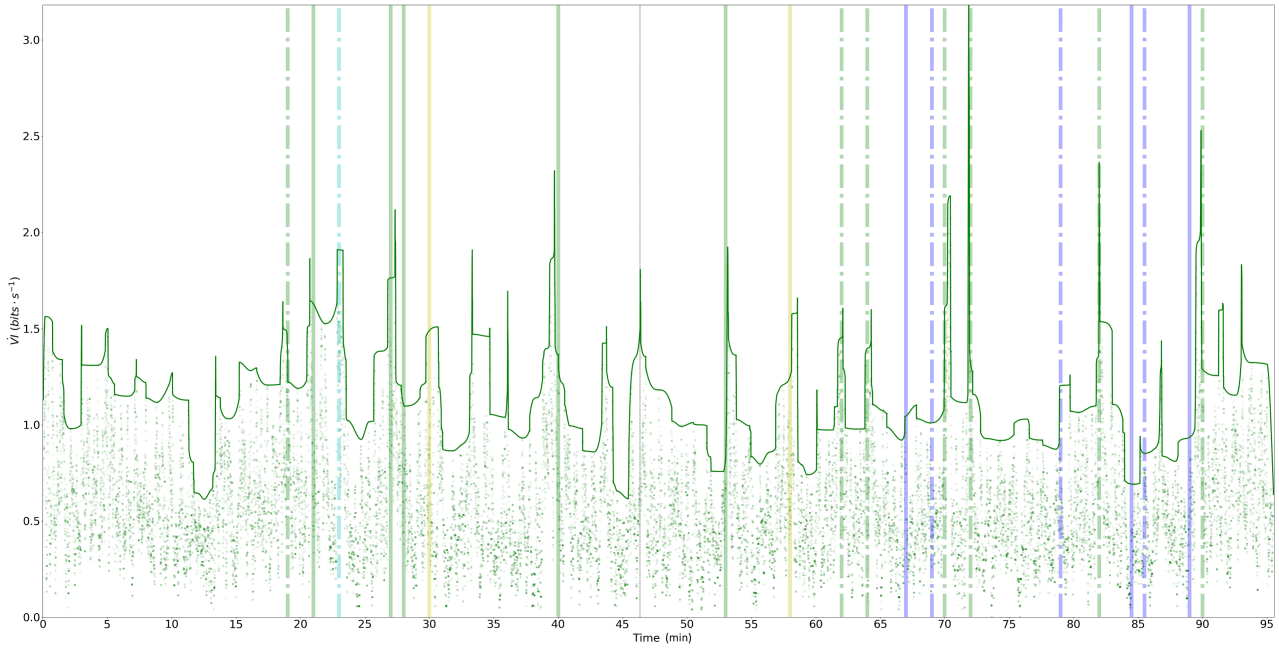
5.4.6 Player's VI contribution for simplex transitions is related to his role

To analyse a player contribution to the overall VI, we apply equation 5.2. His individual $\dot{VI}$, can be compared to the average $\dot{VI}$ per player. This may be useful to assess his activity during

the match (figure 5.4). Beyond the trivial low VI observed for the goalkeepers, we observed anecdotal differences between forward, midfielders and defenders consistent with literature reports (Di Salvo et al., 2007).



(a) Match 1, 0-0



(b) Match 2, 2-1

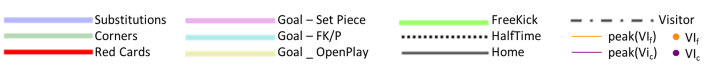


Figure 5.3. Plots for two matches where green points are observations of $\dot{VI}$ at each sample transition, and the colored line the respective peak envelope. $\dot{VI}$ seems to be heavily correlated with match events, such as corners, where a high level of player repositioning is expected, and player substitutions, usually associated with a trough in $\dot{VI}$. It is also visible at minute 92 in 5.3a that the match virtually "stopped" during the send-off of two players from opposing teams.

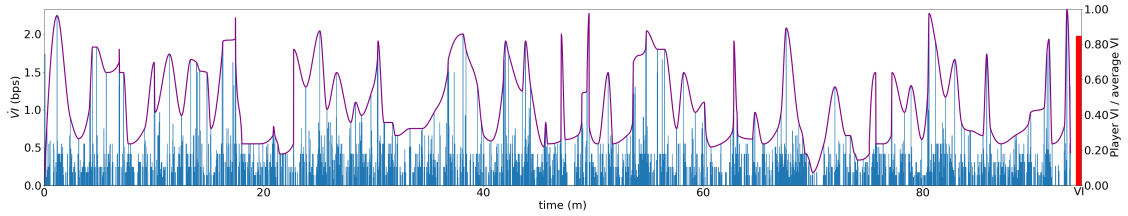


Figure 5.4. $\dot{VI}$ for a single player, in a single match, with maxima envelope. His total $\dot{VI}$ is compared against the match average for the whole match on the red bar on right hand side of this plot. In this case, a center forward player is represented, showing a lower than average $\dot{VI}$, which may be expected, because a forward is typically less active than the other players during his team defensive sub-phases of the match.

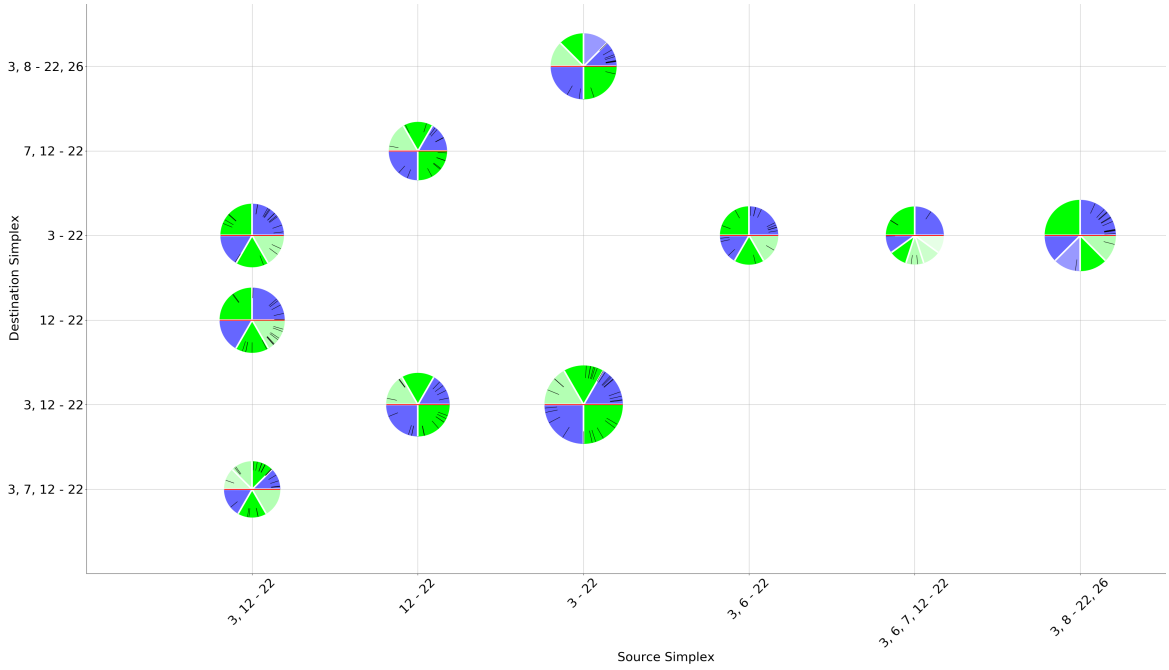


Figure 5.5. This chart shows the top ten simplex transitions player 22 of match 1 (figure 5.3a) was involved in, as well as their formation. His contribution to the match VI resulting from participating in these simplex transitions, is proportionally encoded in the area of the circle: larger circle signifies higher contributions. Each formation is coded in color and shade, with green and blue representing, respectively, home and visitor players, and the number of shades the number of participating players in the simplex. Each tick signals a transition and the match moment when it occurred, with a full match taking a full circle. The lower and upper semicircles describe, respectively, the formation of the prior (source) and immediately subsequent (destination) simplices, where the player was involved. Finally, simplices are identified by the participating players' numbers, with home players first, followed by visitors. Player 22 is a visiting forward, and as seen in the picture, is frequently observed alone (the single shade of blue in the semi circles) in a simplex with opposing back player(s), a typical pattern. Transition from formation $3-22$ to $3,12-22$, when home player 12 joins the simplex, has the highest accumulated VI contribution from player 22. It occurs throughout the match but with an emphasis in the first half of the first 45 min. Player 22 is supported by a teammate in only two transitions out of the 10 represented.

We visualize the type of transition, color coded to denote the number of home and visiting players involved. Each simplex transition plot is scaled by overall VI contribution for that set of transitions, and details when those transitions occurred (see figure 5.6).

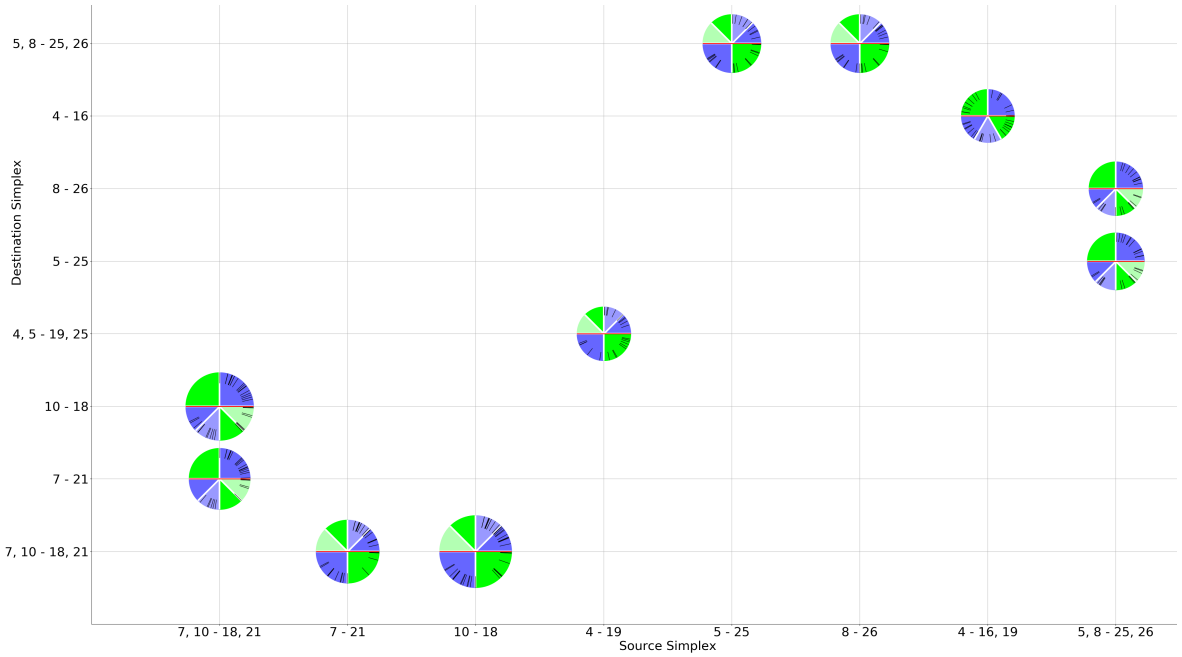


Figure 5.6. This chart uses the same symbolic elements as figure 5.5 but operates at a different level. Each circle represents the overall contribution to the match VI of a whole transition and not just the player's contribution. Here we represent a match top ten transitions. The encoded information in this and in figure 5.5 can be useful to study and train high frequency transitions that contribute significantly to playing dynamics.

5.5 Discussion

A player's performance is dependent on how he perceives and responds to environmental cues that emerge from game play (Travassos, Gonçalves, Marcelino, Monteiro, & Sampaio, 2014). These cues, such as relative positioning of teammates, of adversaries and of the goal, condition the affordances the player has for action, while his own actions change this landscape. This feedback loop generates a complex system that researchers have striven to describe and understand. By using a temporal network of "relationships" to represent the game, we endeavor to uncover insights otherwise missed. The correlations we observed with moments of well known dynamics, endorse our approach. Corners are a prime example, but other events such as free strikes close to the penalty area or interruptions, correlate as well. Other reported observations, such as the impact of fatigue, of the halftime interval or of player role in intensity indicators, were also consistently detected in the 9 matches we analyzed. Although, the study could benefit from a larger sample, the evidence gathered, as shown, is certainly promising.

The proposed way of measuring the soccer game enables a multi-layer decomposition of its dynamics from macro level (a full match) to meso (clusters of players, transitions and teams), to micro (individual players). This enriches the information that can be extracted, helpful to evaluate the dynamics generated by individual players, but also by sets of interacting players, which can uncover which players' structures are more prevalent, how they change and how they impact the overall match dynamics. It is also possible to inspect which simplex transitions a player is involved in, and split his contribution among simplex transitions as shown in figure

5.5. An aggregation of all simplex transition charts provides a full view of a complete match.

As we stated in the introduction, this study is essentially descriptive in nature. This does not mean that the measurements we presented cannot be used for performance analysis. We should however be aware of what the authors in (David & Wilson, 2015) stated: "A greater number of sprints by individuals in a team, amount of ball-related activities, or distance covered had no association with the probability of winning matches". It is true that our method avoids "unproductive" intensity, such as sprints that do not change relative positioning, or, other technical actions that do not increase the agency possibilities a player enjoys. However, given the impact of fatigue, instead of using directly VI as introduced, considering its rate to player and team's work, could intuitively produce a more faithful predictor of performance.

5.6 Final Remarks and Future Research

The presented results endorse the status of $\dot{VI}$ as a measure for game dynamics. The fact that it captures with accuracy and precision well known moments of players jostling for position, supports this interpretation.

With error free and detailed metadata, a more accurate analysis would be possible, especially with concurrent notation hard to capture automatically. The present work is based on prior data, captured and clustered independently, that abstract the reality of a soccer match. Based on the promise shown by the use of information theory and networks as analysis tools, the proposed methods could be valuable to evaluate different approaches to data capture, such as sampling rates, as well as different clustering methods and game representations, such as overlapping, distance weighted networks, non-inertial frames of reference that can accommodate ancillary factors, centroid based clustering, among many others. Extensions to multi-layer networks, where ball action can be integrated, could provide an additional level of insights.

All this is left for future research.

Appendix

To illustrate how $\dot{VI}$ is computed, consider the two moments in a fictional match represented in figure 5.7. The corresponding confusion matrix, which describes the transition of nodes between simplices when going from moment t to t+0.9s during the match, is given in table 5.2. Null matrix elements, as well as unchanged simplices (simplices 1, 2 and 9), do not contribute to informational distance. The contribution of the others is computed according to equation 5.2. The result is shown in table 5.3, where the contribution from each simplex transition can be seen.

The end result is $VI = 0.785615$ or, given that we are measuring a 0.9s interval, $\dot{VI} = \frac{0.785615}{0.9} = 0.872905$ bps.

Simplex	1	2	10	11	12	13	14	15	9
1	2	0	0	0	0	0	0	0	0
2	0	2	0	0	0	0	0	0	0
3	0	0	3	0	0	0	0	0	0
4	0	0	2	0	0	0	0	0	0
5	0	0	0	0	0	0	0	3	0
6	0	0	0	2	1	0	0	0	0
7	0	0	0	0	1	2	0	0	0
8	0	0	0	0	0	0	3	1	0
9	0	0	0	0	0	0	0	0	2

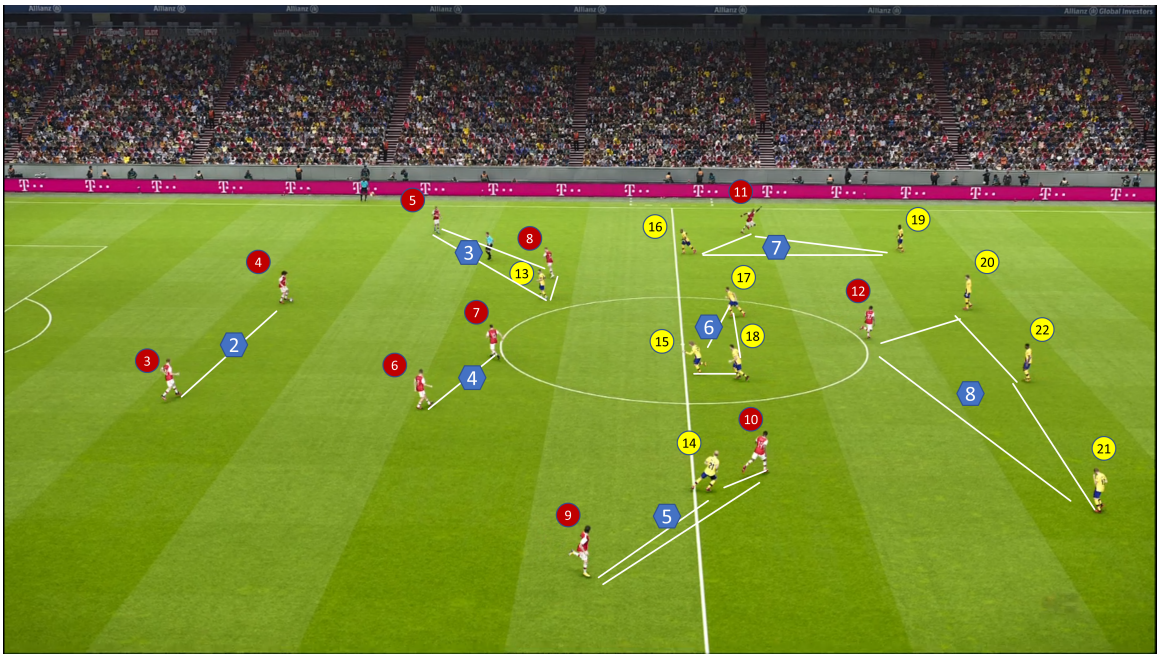
Table 5.2. Confusion matrix going from t to t+0.9s

Simplex	1	2	10	11	12	13	14	15	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0.092121	0	0	0	0	0	0
4	0	0	0.110161	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0.05188	0
6	0	0	0	0.048747	0.107707	0	0	0	0
7	0	0	0	0	0.107707	0.048747	0	0	0
8	0	0	0	0	0	0	0.05188	0.166667	0
9	0	0	0	0	0	0	0	0	0

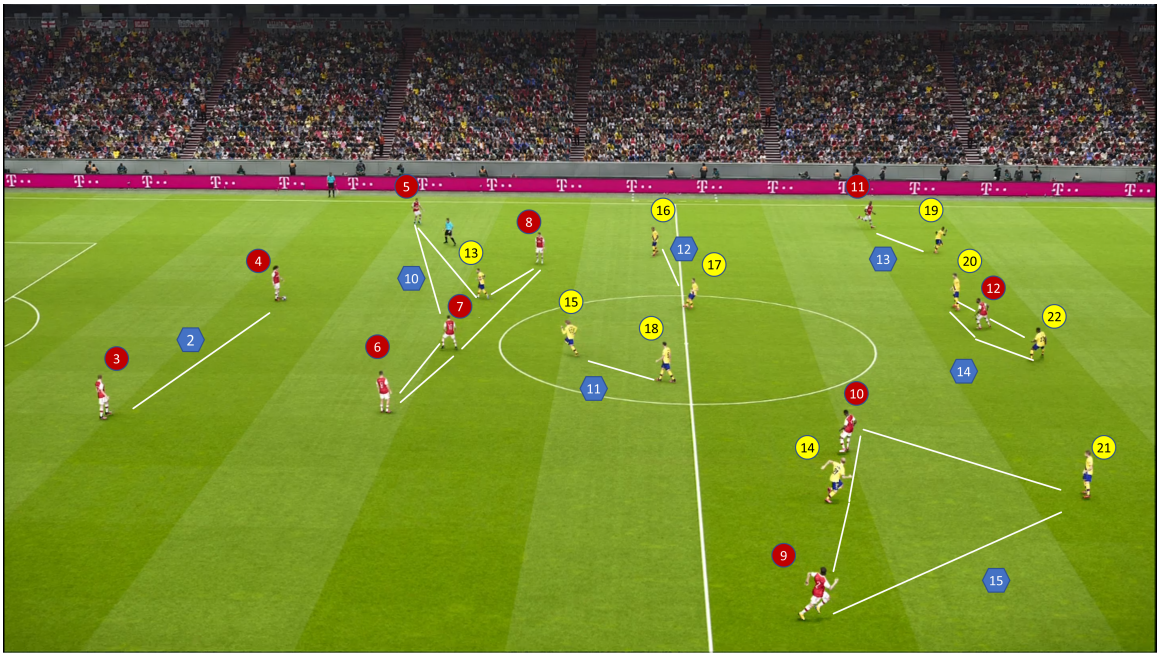
Table 5.3. Computing VI

Match		1	2	3	4	5	6	7	8	9
Result		0-0	2-1	2-2	1-0	3-0	1-0	0-1	2-1	1-0
$\dot{V}I_t$	Avg	0.544	0.591	0.631	0.665	0.622	0.573	0.568	0.599	0.581
	σ	1.255	1.278	1.346	1.369	1.330	1.276	1.273	1.292	1.282
	a	-4.6E-4	-6.0E-4	-2.9E-4	-9.9E-4	1.4E-4	-1.2E-3	-8.7E-4	-1.3E-3	-4.7E-4
$\dot{V}I_h$	Avg	0.277	0.290	0.329	0.330	0.314	0.284	0.301	0.302	0.292
	σ	0.702	0.691	0.774	0.756	0.746	0.696	0.739	0.717	0.711
	a	-6.2E-5	-4.2E-4	2.4E-4	-3.6E-4	-9.4E-5	-6.2E-4	4.4E-4	-6.3E-4	-2.9E-4
$\dot{V}I_v$	Avg	0.267	0.301	0.303	0.335	0.308	0.289	0.267	0.301	0.289
	σ	0.677	0.715	0.718	0.769	0.734	0.712	0.673	0.719	0.709
	a	-4.0E-4	-1.8E-4	-5.3E-4	-6.2E-4	2.4E-4	-6.2E-4	-1.3E-3	-6.6E-4	-1.8E-4

Table 5.4. Average (avg), standard deviation (σ), and linear regression slope (a) for $\dot{VI}$ results (Total, Home and Visitor) for the nine matches used in this chapter



(a) Clustering at time t



(b) Clustering at time t+0.9s

Figure 5.7. Clustering for two moments of a fictional match separated by 900ms. Cluster 1 (goal and goalkeeper of the red team) and Cluster 9 (goal and goalkeeper of the yellow team), are not visible. The clustering process ensures that a node and its closest neighbor are nodes of the same simplex. Home players are numbered in red circles, visitors in yellow. Blue hexagons identify the simplices. White lines are only used to identify simplex membership. Formation for (a) is $\{2^4, 3^4, 4\}$ and for (b) is $\{2^6, 3, 4, 5\}$, which correspond to the row and column sums of the matrix in table 5.2.

Algorithm 3 Clustering players and goals. This pseudo-code describes a non parametric procedure to generate a clustering C, where P is the totally ordered set of all pairs of nodes, sorted by distance. It guarantees that no player is closer to another player, than to the nearest player in its own cluster.

$C \leftarrow \emptyset$
while $P \neq \emptyset$ **do**
    $\{n_1, n_2\} \leftarrow min(P)$ ▷ get the pair with the current shortest inter-distance
    $P \leftarrow P \setminus \{n_1, n_2\}$ ▷ Remove it from P
    $C_f \leftarrow \bigcup C$ ▷ Flatten C
    $n_n \leftarrow |\{n_1, n_2\} \cap C_f|$ ▷ n_n is the number of nodes from the pair in C_f
    if $n_n = 0$ **then** ▷ If none found in C_f
        $C \leftarrow \{\{n_1, n_2\}\} \cup C$ ▷ Create a new cluster
    else if $n_n = 1$ **then** ▷ If only one node found in C_f
        $n_{in}, n_{new} \leftarrow n_1, n_2$ ▷ Assume it is n_1
        if $n_2 \in C_f$ **then**
            $n_{in}, n_{new} \leftarrow n_{new}, n_{in}$ ▷ If not, swap
        end if
        $c \leftarrow \{x \colon n_{in} \in x \colon x \in C\}$ ▷ Get the cluster (c) containing n_{in}
        $c' \leftarrow c \cup \{n_{new}\}$ ▷ Add n_{new}
        $C \leftarrow (C \setminus c) \cup \{c'\}$ ▷ Update cluster in C
    end if ▷ if both nodes found in C_f, do nothing
end while

Chapter 6

Conclusion and future work

I conclude this thesis by briefly discussing its results and the future work they prompt.

6.1 Limitations and discussion

I am of the opinion that a research work is rarely, if ever, complete, and many threads worth exploring were certainly eschewed for the sake of expediency and pragmatism during the work for this thesis. In this section I discuss some of the alternatives we faced when developing the temporal network generator, the classification of community lifecycle events, and the application of our findings and contributions to the study of soccer.

Most of this section is dedicated to explaining the limitations of our work and the usage of various information theoretic measures relevant to this discussion. For clarification we include a visual interpretation of these measures in figure 6.1.1, borrowed and expanded from (T. J. Cover & Thomas, 2006, p.22), that we will be referring to in the following subsections.

6.1.1 Measuring clustering distances

A unique attribute of Syntgen is its capability to evolve a parametric network minimizing stochastic change. The user provides parameters for network size, clustering attributes, node degree distributions and other parameters at every Δ_t, and Syntgen generates a conforming network that is as close as possible to previous network observations, given those parameters. As discussed in the Syntgen article (Pereira et al., 2020), we use the variation of information[1] to measure the distance between network samples, a particular case of measuring the distance of two partitions of the same set. In a nutshell, as explained in figure 6.1.1, the variation of information is the difference between the partitions joint entropy and their mutual information.

[1]Historical note: although the introduction of this measure in the clustering analysis is usually attributed to Marina Meila's 2002 paper "Comparing Clusterings" (Meilă, 2002), although unattributed, it was well known in the information theory community as early as 1991 (T. M. Cover & Thomas, 1991, p.45-46).

Information Theoretic Measures

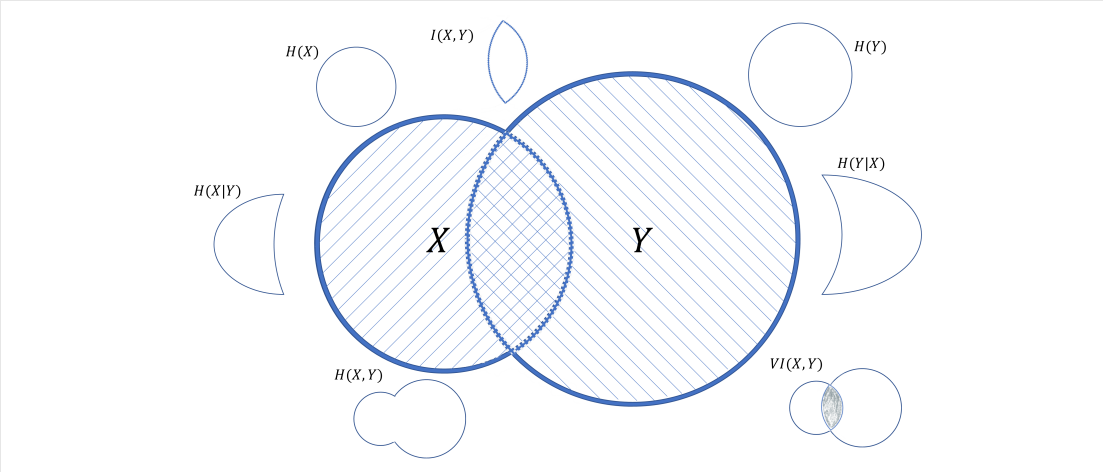


Figure 6.1. Relations of information theoretic measures on clusterings, including information entropy $H(X), H(Y)$, conditional entropy $H(X|Y)$, $H(Y|X)$, joint entropy $H(X,Y)$, mutual information $I(X,Y)$ and the variation of information $VI(X,Y)$.

The information entropy $H(X), H(Y)$ of two clusterings, X and Y, of a set N of nodes, is represented by the areas of their respective blue circles. This entropy varies from a minimum of 0 when the clustering contains a single cluster, denoting total certainty about the location of a node, to a maximum of $\log|X|$ when all clusters in the clustering have the same cardinality, meaning maximum uncertainty of finding the cluster of a given node. The global maximum is $\log|N|$, obtained when partitioning N into $|N|$ clusters. The conditional entropy $H(X|Y)$ of X given Y is represented by the area shaded by the upward diagonals ("the waning crescent moon"). It varies from a minimum of zero when Y is a coarsening of X ("X would be contained in Y"), meaning that the cluster of a node in Y is fully determined by its position in X, to a maximum of its own entropy when $H(Y)$=0. Similarly limits are found for $H(Y|X)$ ("the waxing crescent moon") by swapping Y and X. The joint entropy $H(X,Y)$ is the area circumscribed by the solid blue line and varies from a minimum of $\max(H(X), H(Y))$ when one clustering is a refinement of the other to a maximum of $H(X) + H(Y)$ when the joint probability of finding a node in any given pair of clusters in X and Y is constant. Note that although the entropy of a clustering can go up to $\log|N|$, the global maximum of the joint entropy is also bounded by this same limit, or in other words it is not possible to have a constant joint probability if $H(X) + H(Y) > \log|N|$. The mutual information $I(X,Y)$ is the crossed area delimited by the blue dashed line. It varies from 0 when the joint probability is constant to a maximum of $\min(H(X), H(Y))$ when one clustering is a refinement of the other. Finally the variation of information is the sum of the areas of the "crescent moons", or $H(X|Y) + H(Y|X)$. It can vary from a minimum of 0 when the clusterings are the same to a maximum of $H(X,Y)$ when the joint probability is constant. This has a global maximum of $\min(\log|N|, \log|X| + \log|Y|)$.

Currently, Syntgen always strives to minimize the variation of information, and a core contribution is the analysis of the complexity of this minimization and the development of a heuristic to address it. While the user has indirect means to introduce change points, by dropping or creating nodes as well as change community and connectivity structure, it could be beneficial to remove this limitation and extend the algorithm to allow an adjustable similarity when evolving the network.

Although the general case of comparing partitions of a set was extensively studied in the the last century (Hubert & Arabie, 1985), the emergence of community studies in networks, and the need to benchmark community detection algorithms against a known community structure, i.e. a ground truth, contributed to a renewed interest in this topic from the network science perspective, especially during the first decade of the current century. This is however by no means a settled field, as proposals to address recognized limitations of some of the more popular measures were published in the last few years (Gates, Wood, Hetrick, & Ahn, 2019; M. E. Newman, Cantwell, & Young, 2020). There is however an emerging consensus that a perfect measure does not exist and the specific application domain conditions the choice of the most appropriate one. In this sense, imposing a single method on the user is a limitation of Syntgen, that could however, be easily overcome, by replacing the module used to measuring distances. The justification for using the variation of information was discussed in section 3.3.3, and thus we will not repeat it here. Suffice to say that the mutual information and its variants (Kvålseth, 1987; Strehl & Ghosh, 2003; Yao, 2003), or the adjusted mutual information (Vinh et al., 2010), or any of the methods based on counting pairs of nodes (Fowlkes & Mallows, 1983; Hubert & Arabie, 1985; Mirkin, 1996; Rand, 1971) could have been maximized, instead of minimizing the variation of information.

All of the methods cited in the previous paragraph are concerned with comparing clusterings of the same set of nodes, expanding on the mathematical subject of the partition of a set. Nonetheless, real networks gain and lose nodes as part of their evolution, and this introduction and elimination of nodes is not directly compatible with any of those methods. The information theoretic variables described in 6.1.1 would no longer be applicable as we moved away from partitions of a set to partitions of different sets. This is the domain of network comparison. Broad network comparison is a whole topic onto itself, ranging from isomorphism detection to network distance computation, and a significant volume of topical scholarly research (Tantardini, Ieva, Tajoli, & Piccardi, 2019) can be found. The problem itself is ambiguous. After all, a complex network is a multidimensional object that can be compared along a multitude of attributes, such as meso scale structures like communities or hubs, or micro level such as degree distributions or triads or small motifs (Wills & Meyer, 2020). Most available methods relate to what is usually known as "unknown node-correspondences", where node identification is absent (Soundarajan, Eliassi-Rad, & Gallagher, 2014). This is not the case with the application areas we envisaged. On a temporal network, nodes have persistence and their id is known. We should also recognize that clustering comparison methods do not usually directly take into

consideration the graph edges, although there are exceptions (Poulin & Théberge, 2020). Nevertheless, the emergence of communities is a result of edge topology and thus, indirectly, edges have an influence in clusterings similarity and distance. In contrast, network similarity methods for known node-correspondences are usually directly based on the graphs adjacency matrices, making direct use of the network topology.

The synthetic and empiric networks we studied have a large core of persistent nodes. Intuitively, this is the case with temporal networks. For example, in the hypernetworks we used to represent the soccer game, substitutions and send-offs are the only exception to a fixed set of nodes. In the community lifecycle classification we proposed, new nodes and vanishing nodes are supported, but most of the events relate to dynamics within surviving nodes. In Syntgen we provide a facility to ”kill” a percentage of nodes at any moment, if the user wants to simulate change points. To address the distance problem and apply the variation of information without modifications, we build two extra communities on every time-mediated network pair whose distance we want to compute. In the preceding network sample, a community of “to be born” nodes is added comprised of the nodes of the succeeding network sample. Similarly, a community of “dead nodes” is added in the succeeding network sample, with all the nodes of the preceding community sample that have left the network. This changes the problem of comparing networks to comparing partitions of a set, and the variation of information metric can be used unimpeded. The changes in measurements that this approach introduces are intuitively justified: The variation of information is a local metric, i.e. changes in one cluster only affects measurements where that cluster is involved, so the community pair (new born nodes, dead nodes) do not contribute to the overall measurement as their joint probability is null. Intersections of these extra communities with other communities, may increase the variation of information if their joint probability is not null, but distance increases more when the new born nodes are found in a larger number of communities or likewise for dead nodes. The changes to the measurement are canceled when all the new born nodes and the dead nodes are, respectively, found in, or emanate from, single communities. This is the behavior we would expect when measuring complexity distances, if we consider the death or the birth of a whole community a simpler event than the death or birth of the same number of nodes respectively from or into multiple communities. This was the motivation for the adoption of the method described in this paragraph to compute the information distance of two clusterings of different sets of nodes. However, this is still an open area of research, deserving further investigation, especially for complex systems that may experience extensive sudden changes in its nodes.

6.1.2 Soft clusterings

Another decision taken when developing Syntgen was to simplify some of the attributes that empiric networks exhibit. While we believe the results are still useful, complex systems where elements (nodes) are members of multiple communities are common and those cannot be mod-

eled in the system we developed. Instead of a partition or clustering we usually refer to this arrangement as a cover or soft clustering. Syntgen supports partitions but not covers, that is, nodes can change labellings, understood as their community affiliation, but at any point in time they have a single label. Extending the variation of information to compute differences of soft clusterings is trivial (Meilă, 2002). Consequently, Syntgen can be extended to assign to each node a discrete probability distribution over the set of labels, potentially mutable at every Δ_t, in effect removing this restriction. Similar work has been done for other measures such as the Mutual Information (Esquivel & Rosvall, 2012), but additional research may be needed if, on top of supporting covers, different distance measures are also to be implemented.

6.1.3 Distance normalization

Normalization of the distance measure was also considered throughout our work. Normalized versions of the mutual information are discussed in (Kraskov, Stögbauer, Andrzejak, & Grassberger, 2005b; Vinh et al., 2010), where measures of scale such as $H(X,Y)$, $\max(H(X),H(Y))$, $\min(H(X),H(Y))$, $\frac{H(X)+H(Y)}{2}$ or $\sqrt{H(X)\times H(Y)}$ using the notation introduced in 6.1.1, are used for normalization. The most frequently used measure when comparing clusterings against a known ground truth for algorithm validation, $\frac{2\times I(X,Y)}{H(X)+H(Y)}$, was first proposed in (Kvålseth, 1987) and named in (Fred & Jain, 2003). It can be found in popular network analysis packages, such as Networkx (Hagberg et al., 2008) or Igraph (Csardi & Nepusz, 2006). This measure is however less tight than $\frac{I(X,Y)}{\min(H(X),H(Y))}$.

The authors in (Kraskov et al., 2005b, p.4) also propose, without naming it, a normalization to the variation of information, as $1-\frac{I(X,Y)}{H(X,Y)}$. This however presents the problem of normalizing for the particular instantiation of both clusterings, and not to the space of possible clusterings given a node set. As stated in 6.1.1, the joint entropy of two clusterings can grow up to $\log n$ which may be higher than $H(X,Y)$ of two specific clustering instances.

It is also easy to normalize the variation of information, given that any two clusterings C_1 and C_2 with k_1 and k_2 clusters can only differ up to $\min(\log k_1+\log k_2,\log n)$, and thus a normalized version of the variation of information can be computed as $NVI(C_1,C_2)=\frac{VI(C1,C_2)}{\min(\log k_1+\log k_2,\log n)}$. If the cardinalities of the clusterings is to be considered a confounding variable, but an upper bound $k^*<\sqrt{n}$ for the number of clusters exists, $2\times\log k^*$ can be used as a measure of scale. As a last resort, the number of nodes in the network can also be used, as we know that in all cases $VI\leq\log n$. This upper limit is however less tight. As Meila points out (Meilă, 2007) this does not mean that VI depends on the number of nodes, it just means that with more nodes, more clusterings are possible.

Although it would have been easy to normalize the variation of information, we decided against it, as in many problem domains, such as soccer, the number of nodes is fixed, but the distribution of the number of clusters can vary from moment to moment, and, with it, their joint entropy. In this case the absolute value can be more revealing. On the other hand, if used to

benchmark community detection algorithms against a single ground truth, a normalized value would be more adequate, but that was not our main use case. Future developments may change this requirement, for instance, if Syntgen were to be changed to allow variable similarity when evolving the network, a normalized measure would be easier to manipulate. Just as in the choice of the specific distance measure, here also the problem domain must inform our decisions.

6.1.4 Adjusting for chance

Adjusting measurements for chance has been done for several measures of clustering comparisons (Romano, Vinh, Bailey, & Verspoor, 2016). The objective is to find a null model, that within the constrains of the actual clusterings, measures an unbiased random structure. In simpler terms, if we reshuffle a set of nodes over random clusterings we don't usually get a null similarity (or a maximum distance if comparing for differences). This vanishes at the asymptotic limit in the number of nodes, but can be significant for smaller networks. Usually the adjustment is in the form of $\frac{f(x)-\mathbb{E}(f(x))}{\max(f(x))-\mathbb{E}(f(x))}$. Adjustments to chance were proposed for the Rand Index (Hubert & Arabie, 1985) and the Mutual Information (Vinh et al., 2010). Authors on (Vinh et al., 2010) also propose an adjusted index, which they call the adjusted information distance, using the normalized variation of information introduced in (Kraskov et al., 2005b). This adjustment inherits the complications referred above, and breaks the triangle inequality and thus the metric property of the variation of information. For an alternative, more recent approach, that nevertheless also does not retain the metric property, we refer to (M. E. Newman et al., 2020), where a discussion of the drawbacks of the normalized mutual information (NMI) and proposals for an adjusted metric is found.

When comparing community similarity in our proposed taxonomy, we proposed a Jaccard Index adjusted for chance by using the expected index on a uniform random distribution of nodes over the observed communities. This is justified by the fact that only individual clusters are being compared, and the need to avoid the detection of artifacts of random chance as lifecycle events. This is however an optional feature of our method, that the user can elect to discard. As referred in section 4.4 this adjusted index is compared to an external supplied threshold to guide event classification. This threshold is fixed, but there may be use cases in which this may not be desirable. Another limitation is that our event classification is simply applied to successive network observations, that is, it is memory-less. This may be overly simplistic, as, depending on the underlying complex system, communities may not be continuously active. For instance, authors in (Aynaud et al., 2013) propose what they call a dynamic Jaccard Index for community comparison that is factored by the time distance between community observations. Adopting this or a similar approach would be trivial, but should be guided by the system under study. More work would be needed in the context of real life systems to develop a more flexible approach.

6.1.5 Complex system representation

Representing a complex system as a network has only been briefly touched upon in this thesis. In fact, most of the work here included starts ”downstream”, without questioning how a temporal network can be built to represent a complex system. Even on the application of the theoretical contributions to the game of soccer, the starting point is an already formed network (in this case a hypernetwork) where clusters, or simplices in hypernetwork vocabulary, are formed by player proximity. Other representations are certainly possible, for instance adding weights to nodes to measure the importance a player has in a simplex. We note that extending the variation of information to handle node weights is easy to implement.

When validating the variation of information as a proxy of game dynamics, we used empirical knowledge as the “ground truth” to analyze its correlation to observed measurements. It was assumed that the clustering process did not introduce any errors, which is obviously quite unlikely, how marginal they may be. In many systems the ground truth may be hard to define, and in most social system, it is rarely error-free. This complicates validation. However, if sufficient evidence of a solid ground truth exists, an approach such as the one proposed in this thesis to analyze the soccer game or classifying community events, can be used to validate the clustering process used for representation. The method would involve benchmarking several clustering processes comparing results from our proposed approaches against the ground truth. As an example, in soccer, this could come from expert analysis denoting highly dynamic events, and the ranking of the clustering processes guided by its correlation to the variation of information the network exhibits. A similar exercise would be possible with community lifecycle events.

6.2 Future work

The limitations and discussion presented in section 6.1 prompt many topics for future research. These include extending the temporal network synthetic generator to support community overlaps, tunable temporal similarity, improving the community lifecycle event determination to cater for use cases where a fixed threshold of community survival is too rigid and making use of different distance measurements, including normalization and adjustment to chance.

Normalization was not important for the applications we studied, but it is certainly central to the study of clusterings distances and similarities, and the current state of the art is still an area of active research. As an example, as recently as last year, Mark Newman et al (M. E. Newman et al., 2020) introduced a new measure, they named the ”reduced mutual information” and its normalized version, in an attempt to circumvent some of the difficulties found in the traditional mutual information based measures. It is however not a metric, as it does not obey the triangle inequality rule, nor non-negativity. Although this last property can be useful. Using the Alice and Bob communication archetype, when the normalized reduced mutual information turns negative, it is advantageous for Alice to send Bob the full message instead of leveraging what

Bob already knows about it. This is a threshold that does not exist in the standard measures. In summary, there still seems to be "white space" to further investigate extensions to measurements of similarity or distance of the clusterings of a network, as well as the fitness of the growing number of measures to specific problem domains.

Using a distance measure to evaluate temporal change in clustered networks has shown promise in the study of soccer and could certainly be extended to other complex systems that exhibit clustering and where changes are a topic of interest or concern. For instance, a clustered model of language could be a target for the methods we propose, in order to understand when major changes occur and track their evolution. If change is gradual, understanding the lifecycle of clusters, such as in the evolution of science, their branches and disciplines, could be subject of the methods introduced in 4. This is a well covered topic and the reader is referred to (Fortunato et al., 2018) for an article penned by some of the household names of network science.

The research into the soccer game, described in chapter 5, has a vast range of follow-on topics deserving consideration.

"Upstream" from our work, we can identify several questions regarding the network representation of the game as a complex system. While the existence of well defined clusters of players (defenders/midfielders, goal/goalkeeper, strikers/goals, etc) seems consensual, the actual representation is open to debate. Some properties, like the polyadic nature of player/player/goal interactions, do not generate much controversy, but the role of players in the clusters, the rigid assignment of a player/goal to a single cluster at any point in time, structuring the network representation as a sequence of timeslices instead of a stream, selecting the optimal sampling rate of a time discrete representation, clustering criteria and many others, are all open to discussion and research.

Within the boundaries of our work, several limitations could be addressed. For instance, we have used the distance measure as a proxy for dynamics, and a simple mechanism to assign agency to the individual players. This mechanism, which uniformly distributes the partial VI generated by a cluster transition by the participating players, could be improved by using node weights as discussed in the previous section, as not all players contribute equally to the functioning of a cluster. The soccer ball is absent from the hypernetwork representation of the game we used, and, although there are unique difficulties in directly accounting for it (how to describe a ball in mid-air?), expanding the hypernetwork representation with an additional ball-passing network layer, could accrue useful information revealing other aspects of game play. Additional, time accurate, expertly vetted, notational data could be used to reveal correlations, hitherto hidden, of game patterns with the proposed measures. Extending the analysis to a full season, instead as of a set of matches as accounted in the work for this thesis, would provide a seasonal view, revealing time based patterns, for both team and players.

"Downstream" from our work, an enticing prospect, as mentioned in 5.5 is to factor player's mechanical work into the proposed dynamics measure. Intuitively the factored measure would potentially better correlate with a team's performance.

Finally, the methods, measures and procedures could be applied to other team invasion sports as long as there are enough players to generate a sufficiently large space of viable clusterings. Rugby Union with 15 players per side and Australian Rules Football with 18, would be good candidates. If the space of clusterings is small, such as on small sided soccer games, futsal, beach soccer, or water polo, the granularity of most distance or similarity measures decreases, justifying introducing different representations that can more effectively increase the measure resolution of the game dynamics.

Of the theoretical subjects we covered, one of the most intriguing aspects we came across — but did not fully research — was the decomposition of the informational distance of two clusterings into two summand terms that can be derived from the sizes of their clusters. This sequence of cluster sizes of a clustering is a multiset with an underlying set and associated element multiplicities. For instance a 10 node clustering with 3 clusters, one with 4 nodes, and two with 3 nodes, would have its multiset conventionally represented by $\{4, 3^2\}$. For simplicity, let us refer to these multisets as the function $M(x)$, where the argument x is a clustering. Note the difference: a clustering specifies the clusters the nodes belong to, i.e. their labels, while the function $M(x)$ yields the multiset of its cluster sizes. It is easy to verify that for two clusters X, Y, where $X{=}Y \rightarrow M(X){=}M(Y)$, using the standard equality property of multisets, i.e. elements must coexist in both multisets with the same multiplicity. This is the case when two clusterings have a null informational distance as measured by the variation of information. However, $M(X)$ can be equal to $M(Y)$, even if $X \neq Y$. In fact $VI(X, Y)$ can even be maximal ($\log n$) if $H(X, Y){=}H(X|Y) + H(Y|X)$. Using the representation of figure 6.1.1, $H(X)$ and $H(Y)$ would be equal sized circles with a null intersection, while $X{=}Y$ would be represented by two fully overlapping circles. From these assertions we can see that $M(X){=}M(Y) \rightarrow \min(VI){=}0$ and that $M(X) \neq M(Y) \rightarrow \min(VI) > 0$. Under these conditions, $min(VI)$ is the first term of the variation of information, which we denote as vi_c. The second term, vi_n, is defined as $VI - vi_c$.

Let us illustrate with the minimal example as shown in figure 6.2. Consider four clusterings of a network with four nodes numbered from one to four, $W{=}\{\{1, 2\}, \{3, 4\}\}$, $X{=}\{\{1, 3\}, \{2, 4\}\}$, $Y{=}\{\{1, 2\}, \{3\}, \{4\}\}$ and $Z{=}\{\{1, 3\}, \{2\}, \{4\}\}$, with corresponding multisets of cardinalities $M(W){=}M(X){=}\{2^2\}$, that is with 2 clusters each of 2 nodes and $M(Y){=}M(Z){=}\{2, 1^2\}$, with one cluster of 2 nodes and two clusters of a single node. If we examine the confusion matrices (also known as contingency tables) of W with W $\left(\left[\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right]\right)$, and W with X $\left(\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]\right)$, we see in the first case that there is no uncertainty about the location of nodes in clustering Y, given X, and thus $VI{=}0$. In W compared to X there is total uncertainty about the location of nodes in clustering X, given W, and thus VI is maximal and the mutual information (MI) is null. In this case, $vi_c{=}0$ and $vi_n{=}VI$.

Let us now compare W with Y and W with Z. Their confusion matrices are, respectively, $\left(\left[\begin{smallmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \end{smallmatrix}\right]\right)$, and $\left(\left[\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{smallmatrix}\right]\right)$. As $M(W) \neq M(Y){=}M(Z)$, VI cannot be null and $vi_c > 0$. Inspecting both clustering pairs it is easy to see that the pair W, Y has less uncertainty than the

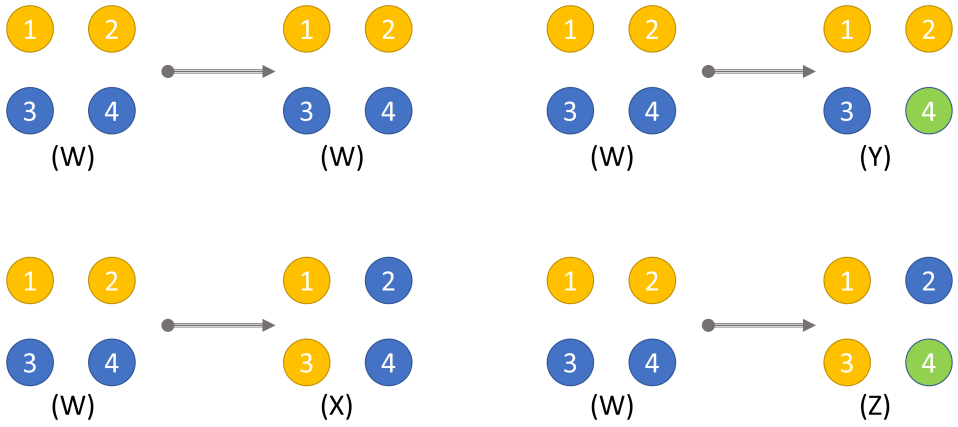
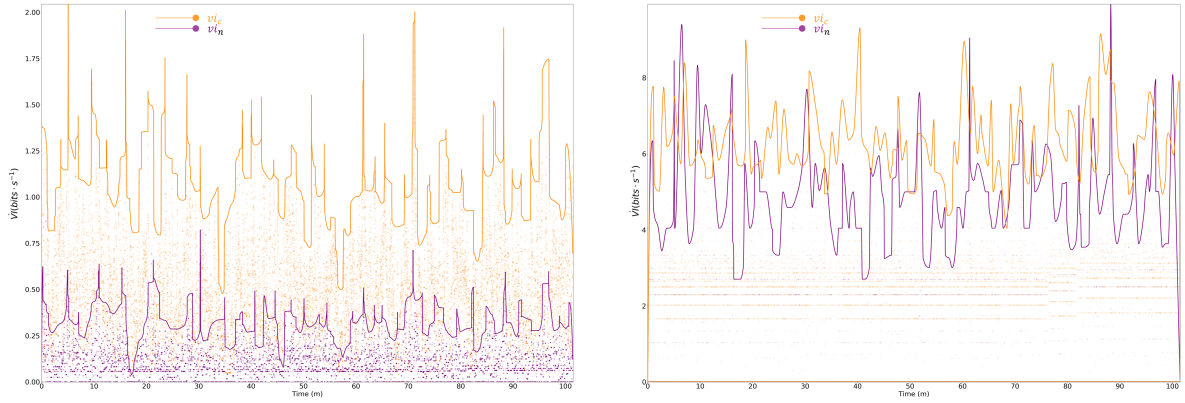


Figure 6.2. Decomposition of the variation of information into vi_c and vi_n as a function of cluster size sequences. 4 transitions between 4 clusterings denoted as W, X, Y and Z. Each circle is a number-identified node. The color of the circle is the node label, denoting its cluster. Clusterings W and X have a multiset of cluster sizes of $\{2^2\}$, while clusterings Y and Z, $\{2, 1^2\}$. Uncertainty is lowest (no uncertainty) going from W to W and $VI{=}vi_c{=}vi_n{=}0$. Uncertainty is highest when going from W to X, with $VI{=}2$, fully accounted by $vi_n{=}2$. Uncertainty going from W to Y ($VI{=}0.5$ $vi_c{=}0.5$) is lower than W to Z, where $VI{=}1.5$ accounted by $vi_c{=}0.5$ and $vi_n{=}1.0$. All VI units in bits.

W, Z, as all nodes in the yellow cluster of W end up in the yellow cluster of Y, while all clusters split when going from W to Z. In fact, for a transition involving clusterings with multisets of cardinalities $\{2^2\}$ and $\{2, 1^2\}$, $VI(W, Y)$ is the minimum, and so $vi_c{=}VI(W, Y)$. For the pair W, Z, as $M(Y){=}M(Z)$, vi_c does not change but, having greater information distance, $vi_n{>}0{=}VI(W, Z) - vi_c$.

In summary it would seem that VI is driven by two terms, one, exacted by the structure of the clusterings, represented by their multiset of cardinalites, and another, resulting from additional "node activity" when switching clusters. The intriguing aspect we referred to above relates to the contextualization of these two terms in real world temporal networks and their associated complex systems. At first glance, those terms do not appear independent, as clusters form from edge connectivity, which is no more than node activity when establishing relations or performing transactions. However, it is not hard to come up with examples of systems where exogenous factors condition the emergence of edges. Spatial systems is such an example. In social systems the likelihood of forming relationships is heavily contingent on close proximity. In designed systems, such as human organizations it is not unusual to encounter subdivisions that condition emergence. When we analyzed the soccer game, we found that when clusters are assembled by physical proximity, cluster distribution is far from random. Quite the contrary, some clusters appear much more frequently than others (J. Ramos et al., 2017). Similar phenomena can be observed if we look at cluster events. In figure 4.5 an example of this distribution can be seen for a single match, where a long tail of events can be observed, while others occur

much more frequently. All of this suggests the influence of design on such systems, and our research question is whether this design is predominant in the partition of the network, while emergent behavior from autonomous node-to-node interaction, that is, that is not imposed by the cluster size sequences, adds to the overall dynamics of the network. In the case of soccer, we would see vi_c as preconceived planning, while vi_n player tactical initiative. This question was not extensively studied, but seems worthy of consideration. The only analysis we did was to compare the envelope of $\max vi_c$ and $\max vi_n$ to the same envelope when averaging over a 5s rolling window. Under the latter conditions, average $\max(vi_c)$ has a much higher contribution to total VI (figure 6.3a), however without a rolling window, the peaks are much closer together (6.3b). If we attribute vi_c to strategy and game design and vi_n to player initiative, we could hypothesize that these results could be justified by critical decision making of best course of action and energy management on the part of the players. This however, needs further study



(a) 4s moving average window **(b)** No moving average

Figure 6.3. On a moving average with sample window of $4s$, $\dot{V}I_c$ has a ≈ 5 times heavier influence on total $\dot{V}I$ than $\dot{V}I_n$ when sampled at 10Hz (6.3a). However, when looking at individual sample maxima, that difference almost disappears (6.3b). If we equate $\dot{V}I$ to energy expenditure, we can interpret this to be due to energy management by players, being judicious about their marking and unmarking efforts.

and is left for future research. Indeed the future research that these results motivate, may not be restricted to team invasion sports, such as soccer, but other domains where external features may condition the cluster structure while leaving some initiative to the individual nodes.

This approach is not dependent on the concrete measure used to measure inter cluster distance or similarity. For instance the Rand index and the mutual information also guarantee non-nullity when the multisets of cluster cardinalities differ. Adjusted to chance, however, any of these measurements may need a different approach to compute the minimum measurement given a pair of clusterings. Let us also recall, as proven in 3.3.3, that the problem of precisely finding vi_c is intractable for anything but networks with few clusters, but for which we proposed an approximate heuristic, easily extendable to other similarity or distance measures. The last item that makes this research thread particular appealing (some would say, risky) is that we have not found any references about it in the network science literature.

Concluding remark

”If you want to have good ideas, you must have many ideas”, Linus Pauling was quoted as saying (Crick, 1995). From the two major threads of this thesis, the theoretical developments and their application to an empiric system, many ideas for future work strands can be spawn. The difficulty is to learn which ones to throw away. Here I made no serious attempt at this task, some may be fruitful, other dead ends. The ones I included are those that I believe are worth the effort of finding out.

Bibliography

Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., & Santoro, N. (2011). On the temporal analysis of scientific network evolution. *Proc. 2011 Int. Conf. Comput. Asp. Soc. Networks, CASoN'11*, 169–174. doi: 10.1109/CASON.2011.6085938

Araújo, D., & Davids, K. (2016). Team Synergies in Sport : Theory and Measures. *Front. Psychol.*, *7*(September), 1–13. doi: 10.3389/fpsyg.2016.01449

Asur, S., Parthasarathy, S., & Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Trans. Knowl. Discov. Data*, *3*(4), 1–36. doi: 10.1145/1631162.1631164

Aynaud, T., Fleury, E., Guillaume, J. L., & Wang, Q. (2013). Communities in evolving networks: Definitions, detection, and analysis techniques. In *Dyn. complex networks* (Vol. 2, pp. 159–200). New York, NY: Birkhäuser. doi: 10.1007/978-1-4614-6729-8_9

Aynaud, T., & Guillaume, J.-L. (2011). Multi-Step Community Detection and Hierarchical Time Segmentation in Evolving Networks. In *Proc. 5th sna-kdd work.* Retrieved from `http://www.complexnetworks.fr/wp-content/uploads/2011/06/snakdd_author.pdf`

Aynaud, T., Guillaume, J.-l., Wang, Q., & Fleury, E. (n.d.). Communities in evolving networks : definitions , detections and analysis techniques. *Comput. Networks*.

Baldoni, V., Berline, N., De Loera, J. A., Dutra, B., Koppe, M., Moreinis, S., & Wu, J. (2014). *A User 's Guide for LattE integrale v1.7.2.* Retrieved from `https://www.math.ucdavis.edu/$\sim$latte/`

Barabási, A.-L. (2015). Network Science: 9. Communities. *Netw. Sci.*.

Barabasi, A.-L. (2016). *Network Science*. Cambridge University Press. Retrieved from `http://networksciencebook.com/`

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science (80-.).*, *286*(5439), 509–512. doi: 10.1126/science.286.5439.509

Barvinok, A., & Pommersheim, J. (1999). An algorithmic theory of lattice points in polyhedra. *New Perspect. Algebr. Comb.*, *38*, 91–147.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third Int. AAAI Conf. Weblogs Soc. Media*, 361–362. doi: 10.1136/qshc.2004.010033

Bazzi, M., Jeub, L. G. S., Arenas, A., Howison, S. D., & Porter, M. A. (2020). A framework

for the construction of generative models for mesoscale structure in multilayer networks. *Phys. Rev. Res.*, *2*(2), 1–32. doi: 10.1103/physrevresearch.2.023100

Behzad, M., & Chartrand, G. (1967). No Graph is Perfect. *Am. Math. Mon.*, *74*(8), 962–963.

Berge, C. (1973). *Graphs and hypergraphs*. Amsterdam: North-Holland. Retrieved from `https://cds.cern.ch/record/105623`

Biagini, F., Brandes, U., Dereich, S., Detering, N., Hunter, D. R., Kauermann, G., ... Wit, E. C. (2019). *Network science* (Vol. 24; F. Biagini, G. Kauermann, & T. Meyer-Brandis, Eds.) (No. 6). Cham, Switzerland: Springer Nature Switzerland. doi: 10.1109/MNET.2010.5634435

Bianconi, G. (2018). *Multilayer networks: Structure and function*. New York, NY: Oxford University Press. doi: 10.1093/oso/9780198753919.001.0001

Bianconi, G., & Barabási, A. L. (2001). Competition and multiscaling in evolving networks. *EPL*, *54*(4), 436–442. doi: 10.1515/9781400841356.361

Blondel, V. D., Guillaume, J.-l., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, *P10008*.

Boguñá, M., Pastor-Satorras, R., & Vespignani, A. (2004). Cut-offs and finite size effects in scale-free networks. *Eur. Phys. J. B*, *38*(2), 205–209. doi: 10.1140/epjb/e2004-00038-8

Bródka, P., Saganowski, S., & Kazienko, P. (2013). GED: the method for group evolution discovery in social networks. *Soc. Netw. Anal. Min.*, *3*(1), 1–14. doi: 10.1007/s13278-012-0058-8

Buldú, J. M., Busquets, J., Martínez, J. H., Herrera-Diestra, J. L., Echegoyen, I., Galeano, J., & Luque, J. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Front. Psychol.*, *9*(OCT), 1–5. doi: 10.3389/fpsyg.2018.01900

Casal, C. A., Maneiro, R., Ardá, T., Losada, J. L., & Rial, A. (2015). Analysis of corner kick success in elite football. *Int. J. Perform. Anal. Sport*, *15*(2), 430–451. doi: 10.1080/24748668.2015.11868805

Castellani, B. (2018). *Map of Complexity Science.* Retrieved from `https://www.art-sciencefactory.com/complexity-map_feb09.html`

Cazabet, R., & Rossetti, G. (2019). Challenges in Community Discovery on Temporal Networks. In P. Holme & J. Saramäki (Eds.), *Temporal netw. theory* (pp. 181–197). Springer, Cham. doi: https://doi.org/10.1007/978-3-030-23495-9_10

Choudum, S. A. (1986). A simple proof of the Erdos-Gallai theorem on graph sequences. *Bull. Aust. Math. Soc.*, *33*(1), 67–70. doi: 10.1017/S0004972700002872

Chykhradze, K., Korshunov, A., Buzun, N., Pastukhov, R., Kuzyurin, N., Turdakov, D., & Kim, H. (2014). Distributed generation of billion-node social graphs with overlapping community structure. *Stud. Comput. Intell.*, *549*, 199–208. doi: 10.1007/978-3-319-05401-8_19

Clauset, A. (2013). The configuration model. *Netw. Anal. Model. CSI 5352, Lect. 11*(October), 1–6. Retrieved from `http://tuvalu.santafe.edu/$\sim$aaronc/courses/5352/`

fall2013/csci5352_2013_L11.pdf

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.*, *51*(4), 661–703. doi: 10.1109/ICPC.2008.18

Clemente, F. M., Martins, F. M. L., & Mendes, R. S. (2016). Analysis of scored and conceded goals by a football team throughout a season: A network analysis. *Kinesiology*, *48*(1), 103–114. doi: 10.26582/k.48.1.5

Cohen, W. C. M. U. (2015). *Enron Dataset.* Retrieved from https://www.cs.cmu.edu/$\sim$enron/

Conway, J. H., & Sloane, N. J. A. (1999). *Sphere Packings, Lattices and Groups* (3rd ed.). New York, NY: Springer. Retrieved from http://www.springerlink.com/index/10.1007/978-3-540-71050-9

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms* (Third ed.). Cambridge, MA: The MIT Press.

Cotta, C., Mora, A. M., Merelo, J. J., & Merelo-Molina, C. (2013). A network analysis of the 2010 FIFA world cup champion team play. *J. Syst. Sci. Complex.*, *26*(1), 21–42. doi: 10.1007/s11424-013-2291-2

Couceiro, M. S., Clemente, F. M., Martins, F. M., & Tenreiro Machado, J. A. (2014). Dynamical stability and predictability of football players: The study of one match. *Entropy*, *16*(2), 645–674. doi: 10.3390/e16020645

Cover, T. J., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Hoboken, NJ, USA: John Wiley and Sons, inc.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory* (1st ed.). New York, NY: John Wiley and Sons, inc.

Cozzo, E., Ferraz de Arruda, G., Rodrigues, F. A., & Moreno, Y. (2018). *Multiplex Networks: Basic formalism and structural properties*.

Crick, F. (1995). *The Impact of Linus Pauling on Molecular Biology.* Retrieved from http://oregonstate.edu/dept/Special_Collections/subpages/ahp/1995symposium/crick.html

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Sy*, 1695. doi: 10.3724/sp.j.1087.2009.02191

Dakiche, N., Benbouzid-Si Tayeb, F., Slimani, Y., & Benatchba, K. (2019). Tracking community evolution in social networks: A survey. *Inf. Process. Manag.*, *56*(3), 1084–1102. Retrieved from https://doi.org/10.1016/j.ipm.2018.03.005 doi: 10.1016/j.ipm.2018.03.005

Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *J. Stat. Mech. Theory Exp.*, *09008*(9), 219–228. doi: 10.1088/1742-5468/2005/09/P09008

Dao, V.-l., Bothorel, C., Lenca, P., Dao, V.-l., Bothorel, C., Lenca, P., & Dao, V.-l. (2017). Community structures evaluation in complex networks : A descriptive approach. In

E. Shmueli, B. Barzel, & R. Puzis (Eds.), *3rd int. winter sch. conf. netw. sci. springer proc. complex.* (pp. 11–19). Cham: Springer.

Darst, R. K., Granell, C., Arenas, A., Gómez, S., Saramäki, J., & Fortunato, S. (2016). Detection of timescales in evolving complex systems. *Nat. Publ. Gr.*(November), 1–8. doi: 10.1038/srep39713

David, G. K., & Wilson, R. S. (2015). Cooperation improves success during intergroup competition: An analysis using data from professional soccer tournaments. *PLoS One*, *10*(8), 1–10. doi: 10.1371/journal.pone.0136503

De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M. A., ... Arenas, A. (2013). Mathematical formulation of multilayer networks. *Phys. Rev. X*, *3*(4), 1–15. doi: 10.1103/PhysRevX.3.041022

Di Salvo, V., Baron, R., Tschan, H., Calderon Montero, F. J., Bachl, N., & Pigozzi, F. (2007). Performance characteristics according to playing position in elite soccer. *Int. J. Sports Med.*, *28*(3), 222–227. doi: 10.1055/s-2006-924294

Dongen, S. V. (2000). Performance Criteria for Graph Clustering and Markov Cluster Experiments. *Methods*.

Dorogovtsev, S. N., & Mendes, J. F. (2002). Evolution of networks. *Adv. Phys.*, *51*(4), 1079–1187. doi: 10.1080/00018730110112519

Duarte, R., Araújo, D., Folgado, H., Esteves, P., Marques, P., & Davids, K. (2013). Capturing complex, non-linear team behaviours during competitive football performance. *J. Syst. Sci. Complex.*, *26*(1), 62–72. doi: 10.1007/s11424-013-2290-3

Duch, J., & Arenas, A. (2005). Community detection in complex networks using extremal optimization. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, *72*(2), 1–4. doi: 10.1103/PhysRevE.72.027104

Eagle, N., & Pentland, A. (2006). Reality mining: Sensing complex social systems. *Pers. Ubiquitous Comput.*, *10*(4), 255–268. doi: 10.1007/s00779-005-0046-3

Enugala, R., Rajamani, L., Ali, K., & Kurapati, S. (2015). Community Detection in Dynamic Social Networks : A Survey. *Int. J. Res. Appl.*, *2*(6), 278–285.

Erdös, P., & Gallai, T. (1960). Gráfok előírt fokú pontokkal. *Mat. Lapok*, *11*, 264–274.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17–60.

Esquivel, A. V., & Rosvall, M. (2012). Comparing network covers using mutual information. *Arxiv [Math-ph]*, *1202.0425v*. Retrieved from `http://arxiv.org/abs/1202.0425`

Euler, L. (1736). Solutio problematis ad geometrian situs pertinentis. *Coment. Acad. Sci. Petropolitanae*, *8*, 128–140. doi: 002433.d/232323

Ferro, A., Villacieros, J., Floría, P., & Graupera, J. L. (2014). Analysis of speed performance in soccer by a playing position and a sports level using a laser system. *J. Hum. Kinet.*, *44*(1), 143–153. doi: 10.2478/hukin-2014-0120

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep.*, *486*(3-5), 75–174. doi:

10.1016/j.physrep.2009.11.002

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci.*, *104*(1), 36–41. doi: 10.1073/pnas.0605965104

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., … Barabási, A. L. (2018). Science of science. *Science (80-.).*, *359*(6379). doi: 10.1126/science.aao0185

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Phys. Rep.*, *659*, 1–44. doi: 10.1016/j.physrep.2016.09.002

Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *J. Am. Stat. Assoc.*, *78*(383), 553–569. doi: 10.1080/01621459.1983.10478008

Fred, A. L., & Jain, A. K. (2003). Robust data clustering. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, *2*(May 2003). doi: 10.1109/cvpr.2003.1211462

Gama, J., Passos, P., Davids, K., Relvas, H., Ribeiro, J., Vaz, V., & Dias, G. (2014). Network analysis and intra-team activity in attacking phases of professional football. *Int. J. Perform. Anal. Sport*, *14*(3), 692–708. doi: 10.1080/24748668.2014.11868752

Gates, A. J., Wood, I. B., Hetrick, W. P., & Ahn, Y. Y. (2019). Element-centric clustering comparison unifies overlaps and hierarchy. *Sci. Rep.*, *9*(1). doi: 10.1038/s41598-019-44892-y

Gelardi, V., Godard, J., Paleressompoulle, D., Claidiere, N., & Barrat, A. (2020). Measuring social networks in primates: Wearable sensors versus direct observations. *Proc. R. Soc. A Math. Phys. Eng. Sci.*, *476*(2236). doi: 10.1098/rspa.2019.0737

Génois, M., & Barrat, A. (2018). Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.*, *7*(1), 1–18. Retrieved from `http://dx.doi.org/10.1140/epjds/s13688-018-0140-1` doi: 10.1140/epjds/s13688-018-0140-1

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, *99*(12), 7821–7826. doi: 10.1073/pnas.122653799

Gómez, S., Díaz-Guilera, A., Gómez-Gardeñes, J., Pérez-Vicente, C. J., Moreno, Y., & Arenas, A. (2013). Diffusion dynamics on multiplex networks. *Phys. Rev. Lett.*, *110*(2), 1–5. doi: 10.1103/PhysRevLett.110.028701

Gordon-Roth, J. (2019). Locke on Personal Identity. In E. N. Zalta (Ed.), *Stanford encycl. philos.* (Spring 201 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2019/entries/locke-personal-identity/`.

Granell, C., Darst, R. K., Arenas, A., Fortunato, S., & Gómez, S. (2015). Benchmark model to assess community structure in evolving networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, *92*(1), 1–11. doi: 10.1103/PhysRevE.92.012805

Greene, D. (2010). Tracking the Evolution of Communities in Dynamic Social Networks. In *2010 int. conf. adv. soc. netw. anal. min.* (pp. 176–183). IEEE.

Grund, T. U. (2012). Network structure and team performance: The case of English Premier League soccer teams. *Soc. Networks*, *34*(4), 682–690. doi: 10.1016/j.socnet.2012.08.004

Grünwald, P., & Vitányi, P. (2008). Shannon Information and Kolmogorov Complexity. *arXiv Prepr. cs/0410002*, 1–54.

Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, *433*, 895–900. doi: 10.1038/nature03286.1.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. *7th Python Sci. Conf. (SciPy 2008)*(SciPy), 11–15.

Hartmann, T., Kappes, A., & Wagner, D. (2016). Clustering evolving networks. In L. Kliemann & P. Sanders (Eds.), *Algorithm eng.* (Vol. 9220 LNCS, pp. 280–329). Springer, Cham. doi: 10.1007/978-3-319-49487-6-9

Hewitt, A., Greenham, G., & Norton, K. (2016). Game style in soccer: What is it and can we quantify it? *Int. J. Perform. Anal. Sport*, *16*(1), 355–372. doi: 10.1080/24748668.2016.11868892

Holme, P. (2015). Modern temporal network theory : A colloquium. *Eur. Phys. J.*, *88*(9), 1–30.

Holme, P., & Saramäki, J. (2012). Temporal networks. *Phys. Rep.*, *519*(3), 97–125. doi: 10.1016/j.physrep.2012.03.001

Holme, P., & Saramäki, J. (2013). *Temporal networks* (Vol. 519; P. Holme & J. Saramäki, Eds.) (No. 3). Berlin Heidelberg: Spreinger-Verlag. doi: 10.1007/978-3-642-36461-7

Holme, P., & Saramäki, J. (2019). *Temporal Network Theory*. doi: 10.1007/978-3-030-23495-9

Hopcroft, J., Khan, O., Kulis, B., & Selman, B. (2004). Tracking Evolving Communities in Large Linked Networks. *Proc. Natl. Acad. Sci. U. S. A.*, *101*, 5249–5253.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *J. Classif.*, *2*(1), 193–218. doi: 10.1007/BF01908075

Jaccard, P. (1912). The distribution of flora in the alpine zone. *New Phytol.*, *11*(2), 37–50. doi: 10.1111/j.1469-8137.1912.tb05611.x

Jdidia, M. B., Robardet, C., & Fleury, É. (2009). Communities detection and the analysis of their dynamics in collaborative networks. *Int. J. Web Based Communities*, *5*(2), 195–211. doi: 10.1504/IJWBC.2009.023965

Jeong, H., Néda, Z., & Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *Europhys. Lett.*, *61*(4), 567–572. doi: 10.1209/epl/i2003-00166-9

Johnson, J. (2010). *Hypernetworks for the Science of Complex Systems*. London: Imperial College Press.

Johnson, J. H. (2016). Hypernetworks: Multidimensional relationships in multilevel systems. *Eur. Phys. J. Spec. Top.*, *225*(6-7), 1037–1052. doi: 10.1140/epjst/e2016-02653-4

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *J. Complex Networks*, *2*(3), 203–271. doi: 10.1093/comnet/cnu016

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.*, *51*(3), 455–500. doi: 10.1137/07070111X

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *Int.*

J. Comput. Math., *2*(1-4), 157–168. doi: 10.1080/00207166808803030

Korf, R. E. (1998). A complete anytime algorithm for number partitioning. *Artif. Intell.*, *106*(2), 181–203. doi: 10.1016/S0004-3702(98)00086-1

Korte, F., Lames, M., Link, D., & Groll, J. (2019). Play-by-play network analysis in football. *Front. Psychol.*, *10*(JULY), 1–10. doi: 10.3389/fpsyg.2019.01738

Kraskov, A., Stögbauer, H., Andrzejak, R. G., & Grassberger, P. (2005a). Hierarchical clustering based on mutual information. *Eur. Lett.*, *70*(2), 278–284. Retrieved from `http://arxiv.org/abs/q-bio/0311039`

Kraskov, A., Stögbauer, H., Andrzejak, R. G., & Grassberger, P. (2005b). Hierarchical clustering using mutual information. *Europhys. Lett.*, *70*(2), 278–284. doi: 10.1209/epl/i2004-10483-y

Kvålseth, T. O. (1987). Entropy and Correlation: Some Comments. *IEEE Trans. Syst. Man Cybern.*, *17*(3), 517–519. doi: 10.1109/TSMC.1987.4309069

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, *80*(5), 1–12. doi: 10.1103/PhysRevE.80.056117

Lancichinetti, A., Fortunato, S., & filippo Radicchi. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E*. doi: 10.1103/PhysRevE.96.052311

Langone, R., Mall, R., & Suykens, J. A. (2015). Clustering data over time using kernel spectral clustering with memory. *2014 IEEE Symp. Comput. Intell. Data Min. (CIDM), Orlando, FL, 2014*(December), 1–8. doi: 10.1109/CIDM.2014.7008141

Latapy, M., Viard, T., & Magnien, C. (2018). Stream graphs and link streams for the modeling of interactions over time. *Soc. Netw. Anal. Min.*, *8*(1), 0. doi: 10.1007/s13278-018-0537-7

Latora, V., Nicosia, V., & Russo, G. (2017). *Complex Networks: principles, methods and applications*. Cambridge, UK: Cambridge University Press. doi: 10.1017/9781316216002

Leskovec, J., Kleinberg, J. O. N., & Faloutsos, C. (2007). Graph Evolution : Densification and Shrinking Diameters. *ACM Trans. Knowl. Discov. Data*, *1*(1), 2–. doi: 10.1145/1217299.1217301

Lewis, T. G. (2009). *Network Science - Theory and Applications*. Hoboken, NJ, USA: John Wiley and Sons, inc.

Loemker, L. E. (1969). *G. W. Leibniz: Philosophical Papers and Letters* (2nd ed., Vol. 53; L. E. Loemker, Ed.) (No. 9). Dordrecht / Boston / london: Kluwer Academic Publishers. doi: 10.1017/CBO9781107415324.004

Loera, J. A. D. (2005). The Many Aspects of Counting Lattice Points in Polytopes. *Math. Semesterberichte*, *52*(2), 175–195.

Lopes, A. M., & Machado, J. A. (2019). Entropy analysis of soccer dynamics. *Entropy*, *21*(2), 3–12. doi: 10.3390/e21020187

Lord, F., Pyne, D. B., Welvaert, M., & Mara, J. K. (2020). Methods of performance analysis

in team invasion sports: A systematic review. *J. Sports Sci.*, *38*(20), 2338–2349. doi: 10.1080/02640414.2020.1785185

Mall, R., Langone, R., & Suykens, J. A. (2015). Netgram: Visualizing communities in evolving networks. *PLoS One*, *10*(9), 1–24. doi: 10.1371/journal.pone.0137502

Martin, C. D., & Porter, M. A. (2012). The extraordinary SVD. *Am. Math. Mon.*, *119*(10), 838–851. doi: 10.4169/amer.math.monthly.119.10.838

Martínez, J. H., Garrido, D., Herrera-Diestra, J. L., Busquets, J., Sevilla-Escoboza, R., & Buldú, J. M. (2020). Spatial and temporal entropies in the Spanish football league: A network science perspective. *Entropy*, *22*(2), 1–17. doi: 10.3390/e22020172

McCulloh, I., & Carley, K. M. (2011). Detecting Change in Longitudinal Social Networks. *J. Soc. Struct.*, *12*(1), 1–37. doi: 10.21307/joss-2019-031

Mclean, S., Salmon, P. M., Gorman, A. D., Stevens, N. J., & Solomon, C. (2018). A social network analysis of the goal scoring passing networks of the 2016 European Football Championships. *Hum. Mov. Sci.*, *57*(July), 400–408. doi: 10.1016/j.humov.2017.10.001

Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Phys. Rev. Lett.*, *107*(23), 1–4. doi: 10.1103/PhysRevLett.107.238701

Meilă, M. (2002). Comparing clusterings. *Tech. Rep. 419*. Retrieved from `https://stat.uw.edu/sites/default/files/files/reports/2002/tr418.pdf` doi: 10.1145/1102351.1102424

Meilă, M. (2003). Comparing clusterings by the variation of information. *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, *2777*, 173–187. doi: 10.1007/978-3-540-45167-9_14

Meilă, M. (2007). Comparing clusterings-an information based distance. *J. Multivar. Anal.*, *98*(5), 873–895. doi: 10.1016/j.jmva.2006.11.013

Menczer, F., Fortunato, S., & Davis, C. A. (2020). *A First Course in Network Science*. doi: 10.1017/9781108653947

Mirkin, B. (1996). *Mathematical Classification and Clustering*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Moura, F. A., Martins, L. E. B., Anido, R. O., Ruffino, P. R. C., Barros, R. M., & Cunha, S. A. (2013). A spectral analysis of team dynamics and tactics in Brazilian football. *J. Sports Sci.*, *31*(14), 1568–1577. doi: 10.1080/02640414.2013.789920

Mukherjee, A., Choudhury, M., Peruani, F., Ganguly, N., & Mitra, B. (2013). *Dynamics On and Of Complex Networks* (Vol. 2). doi: 10.1007/978-0-8176-4751-3

Neuman, E. (1978). Uniform approximation by some Hermite interpolating splines. *J. Comput. Appl. Math.*, *4*(1), 7–9. doi: 10.1016/0771-050X(78)90013-X

Neuman, Y., Israeli, N., Vilenchik, D., & Cohen, Y. (2018). The Adaptive Behavior of a Soccer Team: An Entropy-Based Analysis. *Entropy*, *20*(10), 1–12. doi: 10.3390/e20100758

Newman, M. (2006). Modularity and community structure in networks. *PNAS*, *103*(23), 8577–8582. doi: 10.1017/nws.2015.20

Newman, M. (2018). *Networks* (2nd ed.). Oxford University Press.

Newman, M. E. (2003). Mixing patterns in networks. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, *67*(2), 13. doi: 10.1103/PhysRevE.67.026126

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, *69*(6), 5. doi: 10.1103/PhysRevE.69.066133

Newman, M. E. (2009). The first-mover advantage in scientific publication. *Epl*, *86*(6). doi: 10.1209/0295-5075/86/68001

Newman, M. E., Cantwell, G. T., & Young, J. G. (2020). Improved mutual information measure for clustering, classification, and community detection. *Phys. Rev. E*, *101*(4), 1–11. doi: 10.1103/PhysRevE.101.042304

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, *69*(2 2), 1–15. doi: 10.1103/PhysRevE.69.026113

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2000). Random graphs with arbitrary degree distributions and their applications. *arXiv*. doi: 10.1103/PhysRevE.64.026118

Nguyen, M. V., Kirley, M., & García-Flores, R. (2012). Community evolution in a scientific collaboration network. *2012 IEEE Congr. Evol. Comput. CEC 2012*, 10–15. doi: 10.1109/CEC.2012.6256434

Nicosia, V., Tang, J., Mascolo, C., Musolesi, M., Russo, G., & Latora, V. (2013). Graph metrics for temporal networks. *Underst. Complex Syst.*, 15–40. doi: 10.1007/978-3-642-36461-7-2

Osgnach, C., Poser, S., Bernardini, R., Rinaldo, R., & Di Prampero, P. E. (2010). Energy cost and metabolic power in elite soccer: A new match analysis approach. *Med. Sci. Sports Exerc.*, *42*(1), 170–178. doi: 10.1249/MSS.0b013e3181ae5cfd

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*(1-2), 100–115.

Palla, G., Barabási, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, *446*(7136), 664–667. doi: 10.1038/nature05670

Peel, L., & Clauset, A. (2015). Detecting Change Points in the Large-Scale Structure of Evolving Networks. In *Proc. twenty-ninth aaai conf. artif. intell.* (pp. 2914–2920). Association for the Advancement of Artificial Intelligence.

Pereira, L. R., Lopes, R. J., & Louçã, J. (2020). Syntgen: a system to generate temporal networks with user-specified topology. *J. Complex Networks*, *8*(4). doi: 10.1093/comnet/cnz039

Pereira, L. R., Lopes, R. J., & Louçã, J. (2021). Community identity in a temporal network: A taxonomy proposal. *Ecol. Complex.*, *45*(September 2020). doi: 10.1016/j.ecocom.2020.100904

Pereira, L. R., Lopes, R. J., Louçã, J., Araújo, D., & Ramos, J. (2021). The Soccer Game, bit by bit: An information-theoretic analysis. *Chaos, Solitons and Fractals*, *152*(November).

doi: 10.1016/j.chaos.2021.111356

Pilosof, S., Porter, M. A., Pascual, M., & Kéfi, S. (2017). The multilayer nature of ecological networks. *Nat. Ecol. Evol.*, *1*(4). doi: 10.1038/s41559-017-0101

Poulin, V., & Théberge, F. (2020). Comparing Graph Clusterings : Set partition measures vs . Graph-aware measures. *Trans. Pattern Anal. Mach. Intell.*, *8828*(c). doi: 10.1109/TPAMI.2020.3009862

Pulling, C., Robins, M., & Rixon, T. (2013). Defending corner kicks: Analysis from the English premier league. *Int. J. Perform. Anal. Sport*, *13*(1), 135–148. doi: 10.1080/24748668.2013.11868637

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Paris, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U. S. A.*, *101*(9), 2658–2663. doi: 10.1073/pnas.0400054101

Ramos, J., Lopes, R. J., & Araújo, D. (2018). What's next in complex networks? Capturing the concept of attacking play in invasive team sports. *Sport. Med.*, *48*(1), 17–28.

Ramos, J., Lopes, R. J., Marques, P., & Araújo, D. (2017). Hypernetworks reveal compound variables that capture cooperative and competitive interactions in a soccer match. *Front. Psychol.*, *8*(AUG), 1–12. doi: 10.3389/fpsyg.2017.01379

Ramos, J. P., Lopes, R. J., & Araújo, D. (2020). Interactions between soccer teams reveal both design and emergence: Cooperation, competition and Zipf-Mandelbrot regularity. *Chaos, Solitons and Fractals*, *137*, 1–7. doi: 10.1016/j.chaos.2020.109872

Rampinini, E., Impellizzeri, F. M., Castagna, C., Coutts, A. J., & Wisløff, U. (2009). Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue and competitive level. *J. Sci. Med. Sport*, *12*(1), 227–233. doi: 10.1016/j.jsams.2007.10.002

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.*, *66*(December 1971), 37–41.

Ribeiro, J., Davids, K., Araújo, D., Silva, P., Ramos, J., Lopes, R., & Garganta, J. (2019). The Role of Hypernetworks as a Multilevel Methodology for Modelling and Understanding Dynamics of Team Sports Performance. *Sport. Med.*, *49*(9), 1337–1344. doi: 10.1007/s40279-019-01104-x

Ribeiro, J., Lopes, R., Silva, P., Araújo, D., Barreira, D., Davids, K., ... Garganta, J. (2020). A multilevel hypernetworks approach to capture meso-level synchronisation processes in football. *J. Sports Sci.*, *38*(5), 494–502. doi: 10.1080/02640414.2019.1707399

Ribeiro, J., Silva, P., Duarte, R., Davids, K., & Garganta, J. (2017). Team Sports Performance Analysed Through the Lens of Social Network Theory: Implications for Research and Practice. *Sport. Med.*, *47*(9), 1689–1696. doi: 10.1007/s40279-017-0695-1

Ribeiro, M., Henriques, T., Castro, L., Souto, A., Antunes, L., Costa-Santos, C., & Teixeira, A. (2021). The entropy universe. *Entropy*, *23*(2), 1–35. doi: 10.3390/e23020222

Romano, S., Vinh, N. X., Bailey, J., & Verspoor, K. (2016). Adjusting for chance clustering comparison measures. *J. Mach. Learn. Res.*, *17*, 1–32.

Rossetti, G. (2017). RD YN : graph benchmark handling community dynamics. *J. Complex Networks*(January), 893–912. doi: 10.1093/comnet/cnx016

Rossetti, G., & Cazabet, R. (2018). Community Discovery in Dynamic Networks: a Survey. *ACM Comput. Surv.*, *51*(2), 1–37. Retrieved from `http://arxiv.org/abs/1707.03186` doi: 10.1145/3172867

Salmon, P. M., & McLean, S. (2020). Complexity in the beautiful game: implications for football research and practice. *Sci. Med. Footb.*, *4*(2), 162–167. doi: 10.1080/24733938.2019.1699247

Sampaio, J., & MaçÃs, V. (2012). Measuring tactical behaviour in football. *Int. J. Sports Med.*, *33*(5), 395–401. doi: 10.1055/s-0031-1301320

Sarmento, H., Clemente, F. M., Araújo, D., Davids, K., McRobert, A., & Figueiredo, A. (2018). What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review. *Sport. Med.*, *48*(4), 799–836. doi: 10.1007/s40279-017-0836-6

Sengupta, N., Hamann, M., & Wagner, D. (2017). Benchmark Generator for Dynamic Overlapping Communities in Networks. In *2017 ieee int. conf. data min.* (pp. 415–424). IEEE. doi: 10.1109/ICDM.2017.51

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, *27*(April 1928), 379–423,623–656.

Silva, J. R., Rumpf, M. C., Hertzog, M., Castagna, C., Farooq, A., Girard, O., & Hader, K. (2018). *Acute and Residual Soccer Match-Related Fatigue: A Systematic Review and Meta-analysis* (Vol. 48) (No. 3). Springer International Publishing. doi: 10.1007/s40279-017-0798-8

Solé-Ribalta, A., De Domenico, M., Kouvaris, N. E., Díaz-Guilera, A., Gómez, S., & Arenas, A. (2013). Spectral properties of the Laplacian of multiplex networks. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, *88*(3), 1–6. doi: 10.1103/PhysRevE.88.032807

Soundarajan, S., Eliassi-Rad, T., & Gallagher, B. (2014). A guide to selecting a network similarity method. *SIAM Int. Conf. Data Min. 2014, SDM 2014*, *2*(1), 1037–1045. doi: 10.1137/1.9781611973440.118

Spiliopoulou, M. (2011). Evolution in Social Networks: A survey. In *Soc. netw. data anal.* (pp. 149–175). Boston, MA: Springer. doi: 10.1007/978-1-4419-8462-3

Stanton, I., & Pinar, A. (2011). Prescribed Joint Degree Distribution. *CoRR*, *abs/1103.4*, 1–29. Retrieved from `http://arxiv.org/abs/1103.4875`

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.*, *62*(1), 77–89. doi: 10.1016/S0034-4257(97)00083-7

Strehl, A., & Ghosh, J. (2003). Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, *3*(3), 583–617. doi: 10.1162/153244303321897735

Sun, Y., Tang, J., Pan, L., & Li, J. (2015). Matrix based community evolution events detection in

online social networks. *Proc. - 2015 IEEE Int. Conf. Smart City*, 465–470. doi: 10.1109/SmartCity.2015.114

Takaffoli, M., Sangi, F., Fagnan, J., & Zaiane, O. R. (2010). *A framework for analyzing dynamic social networks.* Retrieved from http://www.researchgate.net/publication/228729658_A_Framework_for_Analyzing_Dynamic_Social_Networks/file/d912f507db68956613.pdf

Takaffoli, M., Sangi, F., Fagnan, J., & Zaïane, O. R. (2011). Community evolution mining in dynamic social networks. *Procedia Soc. Behav. Sci.*, *00*, 48–57.

Tantardini, M., Ieva, F., Tajoli, L., & Piccardi, C. (2019). Comparing methods for comparing networks. *Sci. Rep.*, *9*(1), 1–19. doi: 10.1038/s41598-019-53708-y

The Royal Society. (2010). *The Scientific Century* (Tech. Rep.). London, UK: The Royal Society.

Travassos, B., Gonçalves, B., Marcelino, R., Monteiro, R., & Sampaio, J. (2014). How perceiving additional targets modifies teams' tactical behavior during football small-sided games. *Hum. Mov. Sci.*, *38*, 241–250. doi: 10.1016/j.humov.2014.10.005

Tripathi, A., Venugopalan, S., & West, D. B. (2010). A short constructive proof of the Erdös-Gallai characterization of graphic lists. *Discrete Math.*, *310*(4), 843–844. doi: 10.1016/j.disc.2009.09.023

Vázquez, A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, *67*(5), 15. doi: 10.1103/PhysRevE.67.056104

Vilar, L., Araújo, D., Davids, K., & Bar-Yam, Y. (2013). Science of winning soccer: Emergent pattern-forming dynamics in association football. *J. Syst. Sci. Complex.*, *26*(1), 73–84. doi: 10.1007/s11424-013-2286-z

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, *11*, 2837–2854.

Wagner, S., & Wagner, D. (2007). Comparing Clusterings - An Overview. *KITopen*, *4769*(001907), 1–19. doi: 10.1007/978-3-540-74839-7-12

Watts, D. J., & Strogatz, S. H. (1998). Collective dinamycs of 'small world' networks. *Nature*, *393*(4), 440–442.

Weston, M., Batterham, A. M., Castagna, C., Portas, M. D., Barnes, C., Harley, J., & Lovell, R. J. (2011). Reduction in physical match performance at the start of the second half in elite soccer. *Int. J. Sports Physiol. Perform.*, *6*(2), 174–182. doi: 10.1123/ijspp.6.2.174

Wills, P., & Meyer, F. G. (2020). *Metrics for graph comparison: A practitioner's guide* (Vol. 15) (No. 2). doi: 10.1371/journal.pone.0228728

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: the state of the art and comparative study. *ACM Comput. Surv.*, *45*(4), 1–35. Retrieved from http://arxiv.org/abs/1110.5813 doi: 10.1145/2501654.2501657

Yamamoto, Y., & Yokoyama, K. (2011). Common and unique network dynamics in football games. *PLoS One*, *6*(12), 1–6. doi: 10.1371/journal.pone.0029638

Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A Comparative Analysis of Community Detection Algorithms on Artificial Networks. *Sci. Rep.*, *6*(August), 30750. Retrieved from `http://www.nature.com/articles/srep30750` doi: 10.1038/srep30750

Yao, Y. Y. (2003). Information-Theoretic Measures for Knowledge Discovery and Data Mining. *Entropy Meas. Maximum Entropy Princ. Emerg. Appl.*, *119*, 115–136. doi: 10.1007/978-3-540-36212-8_6