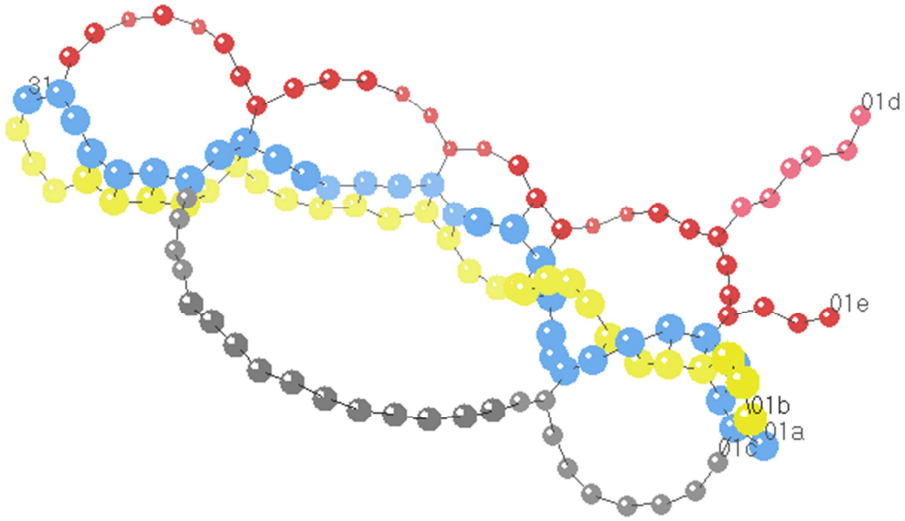
Lausanne COnférence on Sequence Analysis
(LaCOSA*)*
Lausanne 6 – 8 June 2012

Sequence analysis and network analysis:
an attempt to represent and study sequences
by using NetDraw".

By
Ivano Bison



Lausanne 6-8 June 2012

Abstract:

The article presents some alternatives to plotting and analyzing sequences as networks. The point of view proposed is to consider sequence patterns as oriented graphs. Events are treated as nodes in the graph while the arcs, i.e. the ties between events that connect the nodes, are defined by the transitions of subjects between temporally adjacent events. The idea behind the representations discussed is that the patterns combine to form an underlying narrative structure more complex than can be depicted through a network.

The purpose of the paper is to find alternative ways to study sequences. On the one hand, it explores ways different from those of existing visualizations of sequences. On the other hand, it explores alternative ways to extract the information contained in the sequences. The aim in both cases is to find new perspectives from which to study sequences. The overall goal is to open the doors to new ways of identifying potential patterns or underlying structures.

Several graphic examples are provided in the attempt to highlight the limits and potentials of this approach to the study of sequences. The paper uses, for the examples and analysis, data and results of a recent study conducted by myself on class careers in Italy (Bison, 2011b).

Finally, what follows is the result of the first empirical evidence collected by this experimental study, conceived for this conference. I am therefore well aware that what follows will be partial in some of its passages. Moreover, the framework outlined may not be entirely clear and exhaustive in regard to the reasons for using one or the other of the methods presented here.

1.0 Introduction.

What happens if we decide to plot sequences as graphs of a network? Can this approach increase our knowledge of the underlying structure common to the single sequences? Also, will this approach bring to the surface the structures, patterns, and careers still (perhaps) hidden to our eyes and knowledge?

The literature has focused closely on the problems associated with mathematical-statistical techniques to extract underlying patterns from a set of sequences (Abbott and Forrest,1986; Dijkstra W, Taris T, 1995; Billari, 2001; Berchtold and Raftery, 2002; Elzinga, 2003; Elzinga and Liefbroer, 2007; Bison 2006; Ritschard, Gabadinho, Muller, and Studer 2008; Widmer and Ritschard 2009; Gauthier JA, Widmer E.D., Bucher P., Notredame C.,2010; Bison 2011a), but it has devoted little space to the representation and visualization of sequences. This does not mean that little has been done in practice? The tools developed in recent years in Stata (Brzinsky-Fay, Kohler & Luniak, 2006) and R (TraMineR: Gabadinho, Ritschard, Muller, Studer, 2009 & 2011) have made life easier for researchers who want to plot a set of sequences and/or who need to read the content of each cluster obtained from a cluster analysis. What I mean is that, compared with the effort to find tools which allow the treatment of sequences as 'wholes', less attention has been paid to their graphical representation.

At present, there are two main types of representation. The first is based on the visualization of changes in the composition of events over time. Examples are state distribution plots and modal state sequences (Gabadinho, Ritschard, Muller, Studer, 2011). The second type is the sequence index plot (Bison 1999; Bison & Esping-Andersen 2000; Schereer, 2001) which translates a sequence into a coloured line.

These two main ways to represent sequences have a series of limitations. State distribution plots and modal state sequences describe macro changes, but not micro ones. Both display the changes over time of the marginal distributions of the phenomena under observation. These pictures are useful if one is interested in understanding the change over time of the composition of a given phenomenon, for example the proportion of employed workers, unemployed workers and first job-seekers in the labour market.

Unfortunately, these graphs are inadequate if we are interested in understanding the structure of the career pattern. Put differently, state distribution plots and modal state sequences are one-dimensional graphs which furnish little information about micro changes, and even less about how careers develop and the paths followed by individuals and groups.

By contrast, such information is yielded by the sequence index plot. In this case, the level of detail is the most precise and accurate that can be achieved in the visualization of the sequences. It is theoretically[1] possible to represent every single sequence in a single graph. This makes it possible to follow point-to-point, instant-by-instant, the full unfolding of the sequence over time. Moreover, the physical proximity of each line (expertly ordered) to the other produces a second result: that is, insight into the possible existence of common underlying patterns followed by multiple actors.

On the other hand, these are complex graphics and may mislead the researcher.[2] The main problem is that they are interpreted by means of the senses, which sometimes conceal things or show them in a different perspective. For instance, we may be visually attracted by some elements or some graphical structures or specific configurations of colors and shapes in the graph and neglect others. As a result, the graph is read and interpreted in one way rather than another.

Not only sight but also our mind can deceive us. Our mind, in fact, is adept at finding regularities even when they do not exist, or at modifying the complexity degree of the system that we are studying. Hence it is no surprise if we see regularities, patterns or structures that, in fact, do not exist. Similarly, our mind is equally adept at finding differences, irregularities or entropy even in cases where there is order and structure. Obviously I am not here to talk about perception. This I leave to psychologists, who know much, much more than I do. What I want to emphasize is that these graphs may mislead us. These graphs may conceal us a lot more information than is we are reasonably able to capture.

For instance, suppose we want to identify graphically whether there are substantial differences in class careers between men and women. Not being interested in people who do not change class, we select from the database[3] the subset of men and women who, in the first ten years of their working careers, changed class at least once, and who, at the end of the tenth year, were in the urban working class (IIIb+V–VI+VIIa). The data yield (graph 1.0) two sequence index plots depicting respectively the working careers of the Italian women and men who have changed their occupational class position at least once, and who after ten years are in the working class (IIIb+V–VI+VIIa).

The two graphs show some differences between men and women. For example, the proportion of women who start in class IIIa and after ten years end up in Class (IIIb+V–VI+VIIa) is greater than that of men. Conversely, the proportion of men who start in the agricultural classes (IVc and VIIc) is greater than that of women. However, if we exclude the different proportion of classes on the first occupation, and rule out that the different thicknesses of the lines are due to different sample sizes of men and women, other differences are not apparent.

To be precise, the differences found are irrelevant to our purposes. Furthermore, we could obtain this same information with simple frequency distributions and some tables. Also looking with carefully, we cannot in any way argue that at least one specific and distinctive pattern among men or women emerges clearly from the two graphs.

Men and women have the same career patterns. Both groups comprise subjects that, having started their careers among upper classes, entrepreneurs and professionals (I + II), or the middle class (IIIa), or the urban petty bourgeoisie (IVaB), or agricultural petty bourgeoisie (IVc) or the agricultural working class (VIIc), have moved directly to the urban working class (IIIb+V–VI+VIIa).

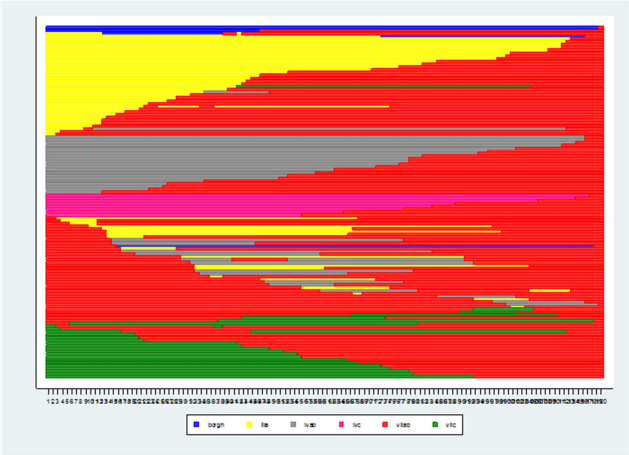
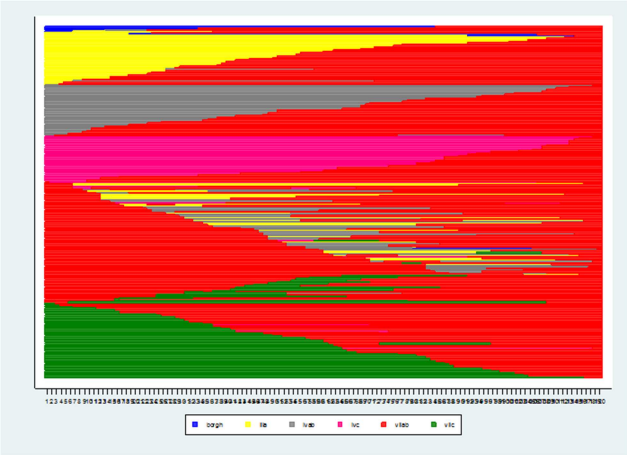
[1] Possible technical limitations may be encountered when the number of sequences is greater than the video resolution and/or the area reserved for the display is smaller than the plot sequences.

[2] Interpretation of these graphs should always be combined with other measures that help the researcher to read the content.

[3] For information about the data see Bison (2011b).

The same complex patterns are to be found in both graphs, i.e. the patterns whereby respondents move across multiple classes in the ten years of observation.

Fig. 1.0. Sequence Index Plot. Working careers that end in class IIIb+V-VI+VIIa.

Women(N=126)	Men (N=242)
	

There are workers (men and women) (a) that start in the middle class, descend to the urban working class, return to the middle class and then finish in the urban working class. Or (b) there are workers who start in the middle class, move laterally to the urban petty bourgeoisie and then end in the urban working class. Finally, there are other workers who (c) begin and end in the urban working class, passing through one or more lower or intermediate classes, such as white-collar middle class, lower middle class, or urban and agriculture working class.

The results from the charts are so clear that they oblige us to draw a single conclusion: that, except for *differences* at the beginning of their careers, already known in the literature, the career patterns of men and women who end in the working class (IIIb+V–VI+VIIa) after ten years are similar. The numbers of the subjects exhibit one or the other pattern changes by gender, but there is no change in either the career pattern or the structure and the concatenation of the temporal events.

The problem, however, is whether this conclusion is actually correct. What if it is wrong? Is, for instance, this conclusion due to our inability to grasp what is depicted by the graph as a ’whole’? Have we failed to understand that these changes of colour from one state to another, apparently similar, describe different patterns for men and women? And, in addition to a common pattern, do there exist others patterns that we cannot see and that are characterized by a different temporal ’shape’ of events? Has the different temporal shape of events generated seemingly equal career patterns but which, in fact, are very different? Finally, is it possible that the two groups are subject to different mechanisms that generate patterns that only apparently are similar but, in fact, are substantially different as regards timing and shape?

We clustered individuals according to the closeness of the shapes of their careers, but we never really considered the ‘*physical*’ form of the career shared by these subjects? Is it a single pattern common to all of them, like a great river that flows slowly to the sea, or is it instead like a myriad of streams that converge to form a large and impetuous river? And finally, how the do events flow?

These are just some of the doubts that have passed through my mind over the years and that have induced me to look elsewhere for new roads, new standpoints from which to view the object ’sequence’.

The use of Social Network Analysis suggested in this paper is double. One, is to find new ways to present results, and second, as a new perspective from which to observe sequences. To do this, however, we need to see sequences, not as individuals moving from one state to another, but as groups of individuals that exhibit common career patterns.

Obviously the main problem is how to bring out this common pattern. To date we have used classification techniques. With this approach we have, in fact, continued to operate from a perspective that differs little from the one normally used with any other type of information given to the variables. We have, in ways of varying complexity, synthesized the information contained in a sequence into some variable which we have then treated in the same way as all the other variables.

In making these changes, however, it has only apparently been possible to capture the entire information structure of a sequence, its shape and its timing. We have assumed that these structures can be captured only at a later time, as a result of some aggregation of a series of sequences within a cluster. Yet there is nothing to guarantee that what we have obtained is real and not the result of some technical mathematical trick. Our mind is very clever at finding regularities even where they do not exist. Furthermore, we ourselves are very able to find models or theories that explain some or other pattern 'ex post'.

Obviously I am exaggerating. In fact, there are reasonable grounds to believe that what we have identified through analysis really exists and is therefore not a mathematical artifact.

What I argue is different. In some ways, in this quest for models that extract information from sequences, we have canceled the sequences themselves. Our focus has shifted from career patterns to the distances between sequences. We have thus lost sight of our object of study, we have hidden behind a number: a distance. We have ceased to observe our research object, the career, as a whole.

Back to observe how careers develop over time. Start again to take into account their dynamic evolution. This proposal starts from these considerations. It is to give physical form to sequences and their underlying generative processes. It is to transform sequences into objects to explore like a DNA chain or to follow like a Google map. And, in the near future, also to be able to model dynamically the processes that have generated a specific pattern, a given sequence.

2.0 Visualizing and studying sequences as networks.

The approach proposed in this paper is not a new one. There are several attempts in the literature to combine sequence analysis and social network analysis. There follows a brief overview of the main applications simultaneously involving networks and sequences.

That networks have entered many fields of science should no longer come as a surprise. A network (or a graph) is a collection of nodes (or vertices), and the connections among them are called arcs, ties, edges. Networks are used to describe, model and analyze an enormous array of phenomena, including physical systems, communication networks, social systems such as networks of friendships or corporate and political hierarchies, physical relationships such as residue interactions in a folded protein, or software systems (Wasserman & Faust 1994, Colizza et al. 2006; Guimera et al. 2007; Kuchaiev et al. 2010). Some of the various applications of networks relate precisely to the study and depiction of sequences.

Although it has been introduced recently, biology is the scientific field where the most developed application of network techniques is made to the study of sequences. Everything suggests that this approach has not only become a useful tool in the study of sequences but in the near future may lead to considerable progress in biology. As Kuchaiev writes, "*Sequence comparison and alignment has had an enormous impact on our understanding of evolution, biology and disease. Comparison and alignment of biological networks will probably have a similar impact.*" (Kuchaiev O., 2010, p.1)

In this regard, there is a wide range of different ways to decline the relationship between sequences and networks. The first and most obvious one is to study of the mechanisms that spread certain viruses in order to discover the sources of infection, as in the case study by Gardy (2011).

> *An outbreak of tuberculosis occurred over a 3-year period in a medium-size community in British Columbia, Canada. The results of mycobacterial interspersed repetitive unit–variable-number tandem-repeat (MIRU-VNTR) genotyping suggested the*

> *outbreak was clonal. Traditional contact tracing did not identify a source. We used whole-genome sequencing and social-network analysis in an effort to describe the outbreak dynamics at a higher resolution.* (Gardy et al., 2011).

Or, from a historical perspective, the evolution of the *human mitochondrial genome* as studied by Herrnstadt et al. 2002.

> *The evolution of the human mitochondrial genome is characterized by the emergence of ethnically distinct lineages or haplogroups. … We have used reduced-median-network approaches to analyze 560 complete European, Asian, and African mtDNA coding-region sequences from unrelated individuals to develop a more complete understanding of sequence diversity both within and between haplogroups.* (Herrnstadt et al. 2002, p.1)

Applications, however, do not only mimic the application of social science to the study of diffusion mechanisms. Another application concerns the study of cells and their relationships. The aim is to redefine the system in which the parts are connected together to form a 'whole'. As Yoon states:

> *… cells are not mere collections of isolated parts. Biological functions are carried out by collaborative efforts of a large number of cellular constituents, and the diverse characteristics of biological systems emerge as a result of complicated interactions among many molecules. As a consequence, the traditional reductionistic approach, which focuses on studying the characteristics of individual molecules and their limited interactions with other molecules, fails to provide a comprehensive picture of living cells. In order to better understand biological systems and their intrinsic complexities, it is essential to study the structure and dynamics of the networks that arise from the complicated interactions among molecules within the cell."*(Yoon 2012, p.1).

In this same work, Yoom identifies three main applications of comparative network analysis to the molecules that compose cells.

> *"… comparative network analysis methods can be broadly divided into three categories: (1) network querying, (2) local network alignment, and (3) global network alignment. In fact, comparative sequence analysis has been shown to be very useful for predicting novel genes and studying the organization of genomes, as well as in many other applications. Similarly, comparative network analysis can serve as a valuable tool for studying biological networks. Comparing the networks of different species provides an effective means of identifying functional modules (e.g., signaling pathways or protein complexes) that are conserved across multiple species, and it can lead to important insights into biological systems."* (Yoon 2012, p.2)

> *"Network querying aims at finding the subnetworks in a "target network" that are similar to a given "query network." This can be used to search for a known functional module or pathway in the biological network of another species, thereby allowing us to transfer the existing knowledge of a well-studied species to other less-studied species. Local network alignment tries to identify similar subnetwork regions that belong to different networks. This method can be useful for detecting novel functional modules that are conserved across different species. Finally, global network alignment aims to find the best overall alignment of two or more networks. This results in a consistent global mapping between nodes that belong to different networks, covering (nearly) all nodes in the given networks."*(Yoon 2012, p.2-3)

Finally, we have what is probably the most important and widespread use of networks to study sequences in biology. This goes by the name of *phylogenetic networks*. In fact, we may say that many of the previous applications fall within this application. These as many of the earlier researches are based on studies of gene mutations.

> *phylogenetic networks should be employed when reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss are believed to be involved, and, even in the absence of such events, phylogenetic networks have a useful role to play* (Huson & Bryant 2006, p.254)
>
> …
>
> *The term phylogenetic network encompasses a number of different concepts, including phylogenetic trees, split networks, reticulate networks, the latter covering both ‘‘hybridization’’ and ‘‘recombination’’ networks, and other types of networks such as ‘‘augmented trees.’’.*
>
> *Recombination networks are closely related to ancestor recombination graphs used in population studies. Split networks can be obtained from character sequences, for example, as a median network, and from distances using the split decomposition or neighbor-net method or from trees as a consensus network or supernetwork. Augmented trees are obtained from phylogenetic trees by inserting additional edges to represent, for example, horizontal gene transfer. Other types of phylogenetic networks include host-parasite phylogenies or haplotype networks.* (Huson & Bryant 2006, p. 255)

In effect, phylogenetic networks are distinguished into three types of representation.

> *Under this very general heading, one can distinguish between a number of different types of networks. Phylogenetic trees constitute one type. A second type is the ‘‘split network,’’ which is obtained as a combinatorial generalization of phylogenetic trees and is designed to represent incompatibilities within and between data sets. A third type, ‘‘reticulate network,’’ represents evolutionary histories in the presence of reticulate events such as hybridization, horizontal gene transfer, or recombination.* (Huson & Bryant 2006, p.254)

The specific aim of the technique is to model graphically the mechanisms of virus recombination (Wain-Hobson, et al. 2003; Huson & Bryant 2006; Bozek, 2009). Or, as argued by Hudson and Bryant (2006)

> *A ‘‘phylogenetic tree’’ is commonly defined as a leaf-labeled tree that represents the evolutionary history of a set of taxa, possibly with branch lengths, either unrooted or rooted.* (Huson & Bryant 2006, p. 254)

It is therefore an attempt to use a graph (network) to represent relationships between the sequences (taxa, allelic profile, etc.). That is:

> *We propose to define a phylogenetic network as ‘‘any’’ network in which taxa are represented by nodes and their evolutionary relationships are represented by edges.* (Huson & Bryant 2006, p.254)

This technique uses a special network representation to model the distance matrix obtained from the deviations between all the sequences.

> *One exception is the use of split networks to visualize distance matrices. The ‘‘phenetic distance’’ between two taxa in a split network is defined as the sum of the weights (or lengths) of the edges along a shortest path between the taxa (Bryant and Moulton 2004). This distance can be computed directly from the associated splits and weights and does not change for different split network representations. The split network, then, is a graphical representation of a collection of splits with weights. The interpretation of the network therefore depends on exactly how the splits were constructed and assigned weights. As we shall see, this varies considerably between methods and between applications.* (Huson & Bryant 2006, p.256)

I think that what the possible applications of this type of technique could be in our field is clear. For example, consider applications in the study of the deviations of a group of sequences

from an ‘ideal type’ sequence as proposed by (Abbott and Hrycak, 1990, Scherer, 2001; Wiggins et al., 2007; Martin et al., 2008). Another application could be to facilitate exploration of the matrix of the distances obtained by optimal matching techniques. Finally, the application that may be the most interesting, at least for the purposes of this article, is the possibility of observing, through the graphical display, the possible existence of distinct career patterns from a given set of sequences.

Not only biology has sought help from network analysis in interpreting relationships between sequences. Even before the development of this interest among biologists, social scientists had begun to test the ground. In particular, historical sociologists in the early 1990s began to use networks to solve their problems in studying sequences of historical events and/or narrative events.

This enabled them to make substantial contributions to the understanding of particular historical problems through the application of network models (Bearman 1993; Gould 1995, 1996; Padgett and Ansell 1993; Rosenthal et al. 1987; Barkey and Van Rossen 1997; Brudner and White 1997; White et al. 1999; Bearman, Faris and Moody 1999; Franzosi (1999), 2004; Bearman and Stovel 2001; Bearman 2002; Bearman, Moody and Faris 2003). Their applications focused on persons, institutions, lineages, and other elements linked by flows of resources, patronage, joint commitment, and kinship (Bearman, Faris and Moody 1999).

Among these scholars, it is mainly Peter Bearman (Bearman, Faris and Moody 1999; Bearman and Stovel 2001; Bearman 2002; Bearman, Moody and Faris 2003) who has developed a new line of inquiry known as *Narrative Network.*

> *By 'narrative network', we refer to a strategy for representing narrative life histories as networks of elements. Applying standard network techniques to this representation of the narrative, we are able to identify some core elements of the process* … (Bearman & Stovel, 2001, p.71)
>
> …
>
> *The central idea is to represent a historical narrative as a graph, and to utilize methods for analysis of graphs to reveal properties of historical event sequences.* (Bearman, 2002, p. 89)

The idea is to translate narrative sequences made up of events and relationships among events into graphs consisting of nodes and links among nodes. However, translating narratives into networks is not straightforward, and it requires a particular set of assumptions by the researcher. The most important of these is the assumption that the stories told by individual respondents follow a set of socially shared rules that underlie the narrative. Otherwise it would be impossible to compare among stories.

> *We are not suggesting that the method we develop here is applicable only to odd stories. By non-canonical we do not mean pathological or disturbed. Canonical expectations are the fundamental expectations that organize our experience. These expectations tend to be extremely simple. One example is the 'continuity principle' that governs our daily experiences. The continuity principle (as canonical expectation) asserts, for example, that when driving, our car will continue to go down the road, that empty spaces behind and in front of us will not suddenly become full of things, and that the road behind us will stay behind us even after we have driven past it, as will the road in front, and so on. It is culturally impossible to tell stories in which all of the elements are organized by canonical expectation; for example, 'I went for a walk and after I put my left foot down ! put my right foot down and then my left foot down and ... And when I turned around the path that I had been walking on was still there' is only a promising beginning of a story if a non-canonical event (the path disappeared, for example) around which a story could be organized is introduced. Abstracted to life stories, there is a more general point. Life stories are accounts of how we became who we are. Once it becomes possible to tell a life-story, the account that is told cannot by definition be canonical, e.g. the life lived is the inexorable byproduct of the canonical way of life, for example, the life of the peasant for whom death is meaningful. Life-stories are consequently a hallmark of modernism*." (Bearman & Stovel, 2001, p.70)

These are not the only attempts made within the social sciences to translate sequences into networks. Again with the intent to use networks to represent narrative sequences, there are those scholars who have set out to describe how individuals use information and communication technology.

> [We use] … *the narrative network as a device for representing patterns of "technology in use." The narrative network offers a novel conceptual vocabulary for the description of information and communication technologies (ICTs) and their relationship to organizational forms. We argue that as ICTs have become increasingly modular and recombinable, so have organizational processes and forms. The narrative network draws on concepts from structuration theory, actor network theory (ANT), and the theory of organizational routines. A narrative network expresses the set of stories (performances) that have been, or could be, generated by combining and recombining fragments of technology in use.* (Pentland and Feldman 2007, p. 781)

There are, finally, attempts that go in the opposite direction, from the network to the sequences. For example, Vedres and Stark (2002) reflected the changing relationships over time detected in a number of companies in states that are in turn encoded as sequences. The latter, in turn, are treated with the standard techniques of sequence analysis. Specifically:

> *Property pathways are conceptualized as the patterned sequences of change that firms undergo 1) in the composition of their ownership structure and 2) in their position within network structures of ties to other enterprises. These career pathways are neither unidirectional nor plotted in advance. The landscape and topography of the socioeconomic field are given shape and repeatedly transformed by the interaction of the multiple strategies of firms attempting to survive in the face of variable political, institutional, and market uncertainties. …*
>
> *To identify patterns of change, the study draws on sequence analysis, a research tool that makes possible the study of historical processes in an eventful way similar to historiography while retaining social scientific abstraction. Whereas sequence analysis has given us a perspective on careers as historical processes but has not been applied to business organizations, network analysis has been applied to business organizations but has not been done historically. The methodological innovation at the heart of this study is to combine the tools of sequence analysis and network analysis to yield a sequence analysis of changing network positions.* (Stark & Vedres, 2002, p.74).

3.0 The narrative: a bridge between sequences analysis and networks analysis.

A sequence is a story. A sequence is a standardized and orderly narrative in which all the elements/events/states that compose it are temporally ordered. As pointed out by Bearman and Stovel (2001):

> *A basic requirement is that our data structures must be longitudinal (though they need not be prospective). That is, they must be capable of revealing process. A more fundamental requirement is that our data reflect the elements that organize the process, as versus those selected from the analyst's hat. One source of data that meet both requirements are life stories. Life stories provide an 'endogenous' account of how authors got from 'there' to where they are. Just like theories, life stories organize facts (elements, states, events, etc.) into interpretable sequences and patterns to reveal a process.* (Bearman & Stovel, 2001, p.76)

Here 'order' means that each element in the sequence is in temporal relation with what precedes it and what follows it.

A sequence is a chain composed of as many rings as there are distinct states in the narrative encoded in it. It is a directed graph, where each event/state is a node and each node is tied to what precedes and what follows it by the temporal line of events.

Finally, many are the reasons that lead us to believe that the advantages outweigh the disadvantages into bringing in the analysis of sequences within the network analysis. The main is because in this way we put in communication two worlds that, in fact, pursue the same purpose, that is to overcome the variable-centric vision in search of common structures.

> *From the view of social network analysis, the social environment can be expressed as patterns or regularities in relationships among interacting units. We will refer to the presence of regular patterns in relationship as* structure. (Wasserman and Faust, 1994, p.3)

The second reason is that in this way we can combine the power of the synthesis of sequences with the power of the representation of the network in order to treat complex structures in a different way.

> *By treating events as nodes and relations between events as arcs, narrative sequences of elements are transformed into networks. By representing complex event sequences as networks, we are able to observe and measure structural features of narratives that may otherwise be difficult to see.* (Bearman, Moody and Faris 2003, p.64)

Adopting the network analysis perspective in the study of sequences is, however, to remain within the fundamental paradigm of sequences analysis, whose ultimate purpose is to identify regularities and structures and to understand the phenomenon as a whole.

> *The general idea is that network representation of narrative sequences provides insight into the social meanings generated within narratives as a whole.* (Bearman & Stovel, 2001, p.70)

To adopt the perspective of network analysis is also to adopt tools to represent complex structures like sequences.

> *By representing complex event sequences as networks, we are easily able to observe and measure structural features of narratives that may other-wise be difficult to see. There are good reasons to think that insight into the structure of narrative processes may be revealed by network methods that provide insight into social structure. One simple reason is that narrative data and 'network data' have many obvious similarities. Specifically, narrative, historical, and network data are locally dense, often cyclic, knotted, and characterized by a redundancy of ties. These similarities suggest that the analysis of narratives and event sequences using network methods may provide a promising avenue for analysis.* (Bearman & Stovel, 2001, p.71)

It is to bring out the hidden structures with the observation. As Gregory Bateson would say (1984), it is to change the perspective. The purpose is to redefine the object by changing the viewpoint from which it is observed. It is to make a different perspective, certainly new, and for this reason, perhaps capable of furnishing new and complementary information about the underlying structures and generative mechanisms.

Adopting a network perspective means shifting the focus from the actors to the events, and to relationships between events generated by the actors. This action may at first seem to contradict the principles that underlie the analysis of sequences. In fact, the primary goal is to study the evolution of the phenomena observed in their individual careers in order to build a "… logical narrative with an inherent telos." (Abbott, 1990, p.141).

However, on the need to shape a common structure, at a certain point, through the classification techniques we leave the individual perspective to examine the underlying structures

representative of groups of sequences. Nor is this shift of focus from the actors to the events new in sequence studies, for it has already been proposed (Billari, Furnkranz, and Prskawetz 2006; Ritschard, Gabadinho, Muller, and Studer 2008). This has yielded new knowledge about typical sequences of states or events. In all these cases, the goal remains the same: to determine how events combine together to form coherent structures of a transition between time t and time t +1 which we denote with the term 'career'.

4.0 From the sequences to the network.

To obtain this new representation, the first step is to define what the nodes are and what the links between nodes are. We can conceive a sequence such as a recording of the succession of states, observed at regular time intervals, on the same unit of survey. For example, the sequences {EUEE} and {EUEU} may be a monthly recording of the employment positions of two subjects in the first four months of their careers.[4]

However, nothing prevents us from representing each of these two sequences as a directed graph, where the nodes are the states observed and the ties between nodes are oriented according to the temporal relationship between the states.

The first sequence {EUEE} would take the following form:

Fig. 1.0, Network plot of sequence {EUEE}

The second sequence {EUEU} would take the following form:

Fig. 2.0, Network plot of sequence {EUEU}

In this depiction, each node is in temporal relation only with the node that precedes it at time t-1 and with the node that follows it at time t+1. In the first sequence, for instance, because the node E, at time t_1, is the first one, it only has a link to U at time t_2. The node U, at time t_2, instead has two links, one incoming from node E which precedes it at time t_1, and one outgoing to E, which follows it at time t_3.

This view, however, adds nothing to our knowledge about the career structure followed by the two actors. We have only a different way to transcribe the information collected.

Yet, on closer inspection, these two sequences have several elements in common. This suggests, for example, that, at least as regards participation in the labour market, these two actors have the first part of their careers in common. Or, in other words, these two workers share elements which constitute a common pattern for the initial parts of their careers.

On reaching this conclusion, we must shift our focus from the individual sequences to what they have in common and what differentiates them. This enables us to configure a new and more complex structure in which the common and the distinct elements are combined to form a new sequence with characteristics different from those that generated it. We have moved from a

[4] These sequences, in fact, are a transcript of the stories that the subjects recount to reconstruct their work histories. Where, for example, for the first sequence {EUEE}, the story is: "I was unemployed after my first month of work, but fortunately the next month I found a new job that I took for two months."

sequence based on the relationship between states to a sequence based on the relationship between events.

In this process of abstraction/generalization, we start from individual sequences and gradually define a new sequence/graph in which the events and relationships between events have been replaced the states and the relations between states. We have thus moved progressively from the individual path to the common pattern: from sequences to the narrative.

At this point also the display and the nature of the graph changes.

The nature of the nodes changes. We move from a list of individual states to a configuration of possible collective events which individuals can access. It is change, for instance, from the class position occupied by the subject at time **t** to the social class at time **t** occupied by one or more actors. Each node of this chain ceases to be an individual condition and becomes a collective condition, an event: for instance, a social class. The weight of this node/class is given by the number of individuals occupying this state at time **t**.

The type of relationship between the elements/nodes of our graph/sequence changes. Previously, the relationship between two elements was temporal. Element E was connected with element U because U followed E temporally. It could be said that the temporal order also determined the relationship between the elements. Now the relation between two elements is given by the probability of observing a transition between two temporally adjacent states. Thus element E, at time t_1, will be connected with element U, at time t_2, if and only if at least one transition is observed between E and U. In other words, there will be a tie between E and U only if between time t_1 and time t_2 there is at least one subject that moves from being employed to unemployed. The amount of these passages defines the transition probabilities between two events/nodes.

Finally, also the shape of the graph changes. Previously, the only form that the graphs could assume was that of a chain. In this new form, several nodes can coexist in the same unit of time and several relations (links) between temporally contiguous nodes may be established.

This new graph has characteristics entirely different from previous ones:

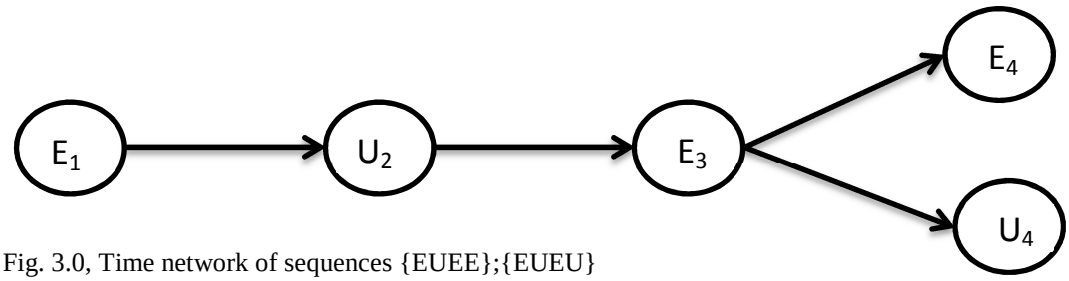


Fig. 3.0, Time network of sequences {EUEE};{EUEU}

(a) Has as many nodes as there are distinct elements in the sequences. Each node represents one event observed on one or more actors at time t. The absence of a node at time t indicates that no one, at that moment, holds that position.

(b) While keeping the temporal order fixed, the link between events is given by transition probabilities. That is, the link is defined as the proportion of individuals who are moving from the node at time t to the node at time t +1. The absence of the link indicates that transitions between the two nodes are not observed;

(c) Multiple nodes can be defined in the same unit of time. In the example of the two previous sequences, observed at time t4 is both an event of unemployment and one of employment.

(d) Several ties can start from the same node or reach the same node. In the case just mentioned, after a common career path, one of the two actors continues to be employed while the other enters unemployment, thus creating a fork in the main pattern.

Unfortunately, we have to deal in our surveys with large numbers of individual sequences composed of dozens of states. This makes it impossible manually to perform the analyses and graphs just introduced. There are several programs that enable the production of networks. The most popular of them among sociologists is NetDraw[5]. NetDraw is social network visualization software[6] which enables the graphic representation of networks including relations and attributes. It has some analytic capabilities and is distributed alone or with UCINET[7].

5.0 From NetWork to NetDraw.

NetDraw, as said, is a tool for visualizing social networks. Its main function is to design networks composed of actors, which are treated as nodes, and links between nodes, which represent the relationships between the actors.

This software is relatively complex and requires a special organization of the input data. In our case, to remedy these problems it is preferred to use NetDraw in combination with UCINET.

To illustrate the preparation of the data, I shall use the sequences in Table 1. Each of the six sequences describes the trajectories followed by six actors through five conditions coded with the letters from (a) to (e). The first sequence, for instance, begins with the state {a} and continues with the states {b – a – b – a – c}.

Tab.1.0. List of six sequences of length six representing six hypothetical class careers.

id	t1	t2	t3	t4	t5	t6
1	a	b	a	b	a	c
2	b	b	a	b	a	c
3	c	b	a	b	b	c
4	c	c	a	b	b	c
5	d	c	a	b	c	c
6	e	e	a	b	c	c

The first step in preparing the data is to recode the states of the sequences in order to ensure that the temporal order of events does not change in the next phases. The strategy is simply to generate a new coding that combines each individual state recorded in sequence with its temporal position. In the example, the new coding of the first sequence will be: {01a, 02b, 03a, 04b, 05a, 06c}. This encoding allows, during visualization and analysis with NetDraw, to identify the individual states/events according to the temporal order in which they have occurred.

At this point, the switch into a directed graph is immediately possible. The proximity of the states, in fact, determines which elements are related to each other, and the temporal order gives the orientation of the arc. For example, in the first sequence, the states 01a and 02b are consecutive, which suggests that the actor (1) has moved from state (a) to state (b) between time 01 and time 02. In graphic terms, this transition is made visually by an arc between 01a and 02b with the arrow pointing to 02b.

By definition, each node $\mathbf{k_t}$ may be connected only with the nodes $\mathbf{k_{t-1}}$ that are located at time **t**-1 and with $\mathbf{k_{t+1}}$ nodes that are located at time **t**+1. Not allowed are ties between states located at distances more than time **t** ± 1. On applying this rule to the first sequence, we have that there is no direct arc between 01a and 03a. Nevertheless, there is still an '*indirect*' pattern connecting 01a and 03a through 02b.

[5] The NetDraw website is: http://sites.google.com/site/netdrawsoftware/

[6] For a quick introduction see Hanneman & Riddle (2005) website http://faculty.ucr.edu/~hanneman/nettext/C4_netdraw.html

[7] This is the main software for social network analysis. Allows the computational aspects of graphs and calculates an ample amount of network measures (Hanneman & Riddle, 2005).

As already mentioned, if we now proceed to construction of the graph, we obtain as many graphs as there are sequences taken individually. Each graph takes the form of a chain and does not add any new information to what we already know.

The solution is to move our focus from the individual sequences composed of sequences of states to events, and to transform the pattern traced by the actors in their movements between states into the links between events. If we consider the six sequences as a matrix of rows and columns, we can see that:

a) Each column of the matrix can be interpreted as a variable whose modalities are the events observed in the sample/population at that time as a repeated cross-sectional sample. At time t_1, for example, only five distinct events are detected in the six sequences because subjects 3 and 4 both begin in position (c). At time t_2 the events are reduced to three. Subjects 1,2,3 are in state (b), subjects in 4,5 are in (c), and subject 6 is in state (e). At time t_3 the number of events is further reduced. All six subjects assume state (a).

b) On observing the phenomenon from another perspective, one notes that between time t_1 and time t_2, the first subject changes from state (a) to (b), the second subject remains in the same state (b). In the case of subjects 3 and 4, who had both begun at position (c), one changes state and goes to (b) while the other remains in (c).

These two different perspectives can be displayed together so that one can see the sequences through a square matrix, called the Adjacency Matrix or Sociomatrix, which has as many rows and columns as there are states (nodes). Each cell (Table 2.0) of the matrix defines the link between two nodes. Cell values greater than zero indicate that there is a link between two nodes, while zero indicates the absence of a relationship. For example, in Table 2, the cell (01a, 02b) has value 1, which establishes the existence of a link between node 01a and node 02b.

Tab.2.0. Adjacency matrix of the six sequences in Table 1.

	01a	01b	01c	01d	01e	02b	02c	02e	03a	04b	05a	05b	05c	06c
01a	0	0	0	0	0	1	0	0	0	0	0	0	0	0
01b	0	0	0	0	0	1	0	0	0	0	0	0	0	0
01c	0	0	0	0	0	1	1	0	0	0	0	0	0	0
01d	0	0	0	0	0	0	1	0	0	0	0	0	0	0
01e	0	0	0	0	0	0	0	1	0	0	0	0	0	0
02b	0	0	0	0	0	0	0	0	3	0	0	0	0	0
02c	0	0	0	0	0	0	0	0	2	0	0	0	0	0
02e	0	0	0	0	0	0	0	0	1	0	0	0	0	0
03a	0	0	0	0	0	0	0	0	0	6	0	0	0	0
04b	0	0	0	0	0	0	0	0	0	0	2	2	2	0
05a	0	0	0	0	0	0	0	0	0	0	0	0	0	2
05b	0	0	0	0	0	0	0	0	0	0	0	0	0	2
05c	0	0	0	0	0	0	0	0	0	0	0	0	0	2
06c	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The directionality of the link is established by the position of the cells compared to the main diagonal. If the cell (with non-zero value) is above the main diagonal, the direction of the link is from the node row to that of the column; if it is below the diagonal, it runs from the node of the column to that of the row. In our example, the cell (01a, 02b) indicates that the link between the two nodes ranges from 01a to 02b.

The cell values represent the force or the weight of the link between two nodes. In our example, the cell (03a, 04b) has a value of 6. This value indicates that between time t_3 and time t_4 six transitions were observed between (a) and (b). In other words, six actors in this range moved from state (a) to state (b), or more precisely, from event (a) to event[8] (b).

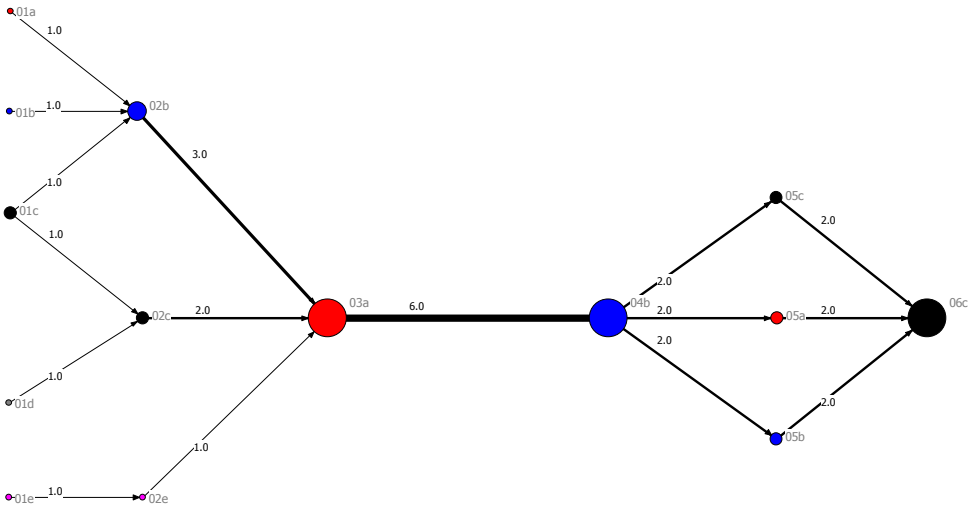
[8] In this new perspective nodes are no longer states within a single sequence but become events experienced by one or more actors at time t.

This particular matrix is now ready for use by NetDraw to produce the graph (fig. 4.0).

Although the input data are quite difficult, NetDraw offers a wide range of possibilities to manipulate the graph. There are three principal possibilities. The first is that the program has a number of optimization routines in the spatial projection of the graph. In particular, NetDraw provides a set of Multi-Dimensional Scale algorithms that automatically place nodes and links in the space so as to maximize the network's readability. The user can still manually intervene in the spatial arrangement of the nodes. The program allows the user to drag, move and change the location of the various components of the graph.

The second characteristic is that weights can be given to the links. In our case, for example, the weight is given by the number of actors that pass through the pattern. NetDraw uses this information on the different numbers by varying the thickness of the link-pattern between the two nodes (Fig. 4.0). A second opportunity is being able to view only the links that exceed a certain weight. This utility program allows the user to create graphs in which links appear only above a certain threshold. This reduces complexity and improves the readability of the network.

Fig. 4.0. Time sequence network of the six sequences in Table 1.0.



The third and most important opportunity offered by NetDraw is that of attaching specific attributes to each node of the network. In our case, we can use aggregate information. For example, we can define an attribute that count the number of actors who stay in the node at time t; or define an attribute 'event' that encodes the type of generic event belonging to each node. For example, nodes 01a, 03a, 05a, which belong to event (a), are associated with code 1, nodes 01b, 02b, 04b, 05b belonging to event (b) are associated with code 2, and so on for the remaining nodes.

NetDraw displays these attributes by transforming them into the colour, size and shape of the node. In the example of Figure 4.0, a colour is given to each node depending on the generic event to which it belongs, while the size is given according to the number of actors in the node at time t.

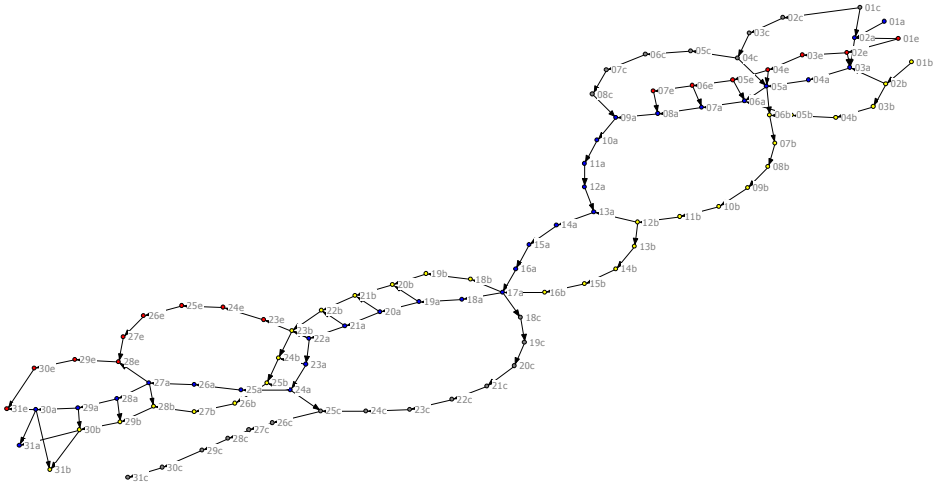
Colour, size and shape can be modified simultaneously on the same node. This makes it possible considerably to increase the network's readability by highlighting structures that might otherwise escape.

Obviously, the graph of Figure 4.0 is based on 'ad hoc' sequences. A real example is shown in Figure 5.0. The network displays a class career pattern identified by a study of the effects of education and social origin on the class careers of a sample of Italians (Bison, 2011b). Specifically, the network of Figure 5.0 shows the careers of individuals that remained mainly in class I-II across the ten years of observation.

To simplify the view it was decided to reduce the observation points from monthly to quarterly. It was reduced by 120 to 31 observations points. The labels of the nodes follow the

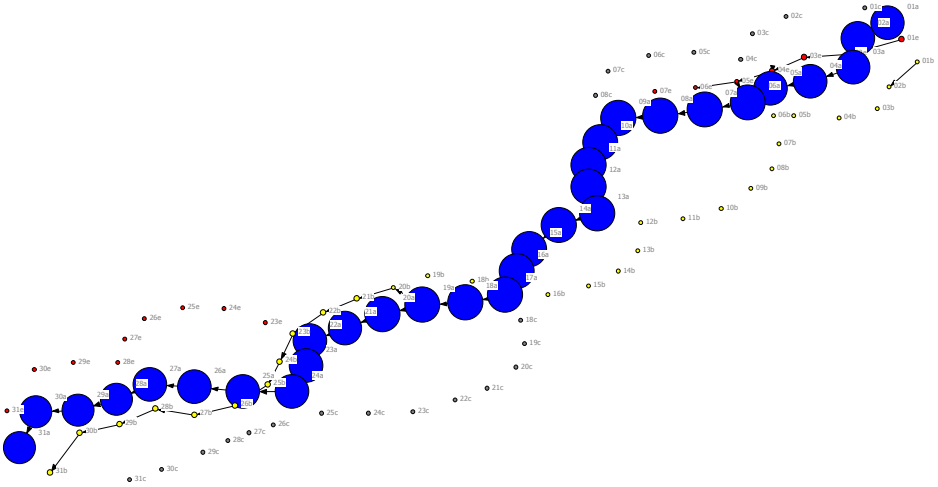
encoding presented above, so the node 01 indicates an event that occurred in the first month of observation, node 02 indicates an event that occurred in the fourth month. The six positions of the EGP schema class are identified by the letters of the alphabet, as follows: (a) I-II, (b) IIIa, (c) IVab; (d) IVc, (e) VIIab; (f) VIIc. Hence the pattern between nodes 17a and 18c shows that at least one actor passes from class I-II to class IVab between the sixty-fourth and sixty-eight month.

Fig.5.0. Time sequence network of cluster (a) I – II



In effect, to be a graph that represents a group of subjects, mainly immobile, it is rather intricate. We would have expected a single chain that a multiplicity of patterns connected. It is clear that this type of representation tends to accentuate even the smallest differences, since also just one person may generate a pattern distinct from others. A solution to this problem is to establish a minimum threshold below which not shown the ties. For example, the graph in Figure 5.1 is the same as Figure 5.0 except that in the latter it has been decided to display only the ties with weights equal to or greater than three: that is, only those patterns with which to plot paths for at least three people. It has also been decided to resize the nodes according to the number of persons passing through the node.

Fig.5.1. Time sequence network of cluster (a) I – II with nodes and ties weighted.



What remains from the previous graph is a main pattern composed of subjects in class I-II and two smaller patterns. The first of these two small patterns consists of workers that enter class I-II after spending at least two years in the urban working class VIIab. The second pattern consists of workers who, after spending more than half of their careers in class I-II, fall in class IIIa. Although

these patterns are residual, and it is sufficient to raise the visualization threshold to six for even these disappear.

The decision to raise the view threshold is justified by the fact that the pattern has a clear structure. However, this involves an arbitrary decision to 'erase' part of the network. What this operation involves, and what the consequences are, is an open question. We shall return to this point later when other examples of application are discussed.

6.0 Some measures to facilitate reading of the network.

It is clear that the simple display of the pattern is not enough to provide information with which to analyze the network fully. In particular, the simple display is not enough to capture the structure of the relationships between events visually and how it changes over time (along the graph). For example, it may be useful to highlight nodes where the largest number of ties converges, as in the case of nodes 02b, 02c and 06c in Figure 4.0. This would show that, in that interval, there is a simplification of the network's complexity and a local reduction in the system's entropy.

It may also be of interest to highlight the nodes from which several ties depart. This would indicate that actors in the same condition at time **t** have taken different paths at time t +1, thus increasing the network's entropy. An example is node 04b, from which three distinct patterns branch out to 05a, 05b and 05c.

Therefore required are specific measures that can be used as attributes of the graph. These measures are displayed through forms, colours and sizes of the nodes in the network, making it graphically evident points, areas or sections of a graph that differ from the rest.

Of course, these indices are experimental and further testing and further development is required. They are mostly mere stylistic exercises that a real and proper definitive measure to be adopted in the analysis of the network.

In order to illustrate the functioning of these indexes, I shall use the network time sequence obtained as a representation of cluster (a) (Bison 2011b) and already used (Fig. 5.0) in the previous section.

A first group of indexes measures the proportion of subjects in each node compared to the total numbers of subjects.

The first is the **Node Probability index** (Figure 6.1a), $NP_{i,t}$,which measures the proportion of subjects **f** that occupy node **i** at time **t** on the total number of subjects **F** in the network .

$$NP_{i,t} = \frac{f_{i,t}}{F}$$

The second and the third indexes measure the deviation of the observed frequency from that expected under the hypothesis of equiprobability of each node. In particular, the **General equal probability index** (Fig. 6.2a) is the natural logarithm of the ratio between the observed frequencies **f** and the average number of expected frequencies under the assumption of equiprobability, where **F** is the total number of subjects, and **K** is the total number of possible events/distinct nodes.

$$GEP_{i,t} = ln\left(\frac{f_{i,t}}{\frac{F}{K}}\right)$$

The index takes value 0 when the observed frequency and the expected frequency are equal, and it assumes positive values if the observed frequency is higher than expected and negative ones if the observed values are lower than expected. The choice of the logarithm is determined by the

need to have comparable quantities both if the numerator is greater and if it is less than the denominator. Graphically (Fig. 6.2a), the size and the sign of the index are defined respectively by the size and shape of the node: The circle shows that the value is positive or zero; the triangle shows that the value is negative. The larger is the size of the node, the greater the deviation from 0.

Fig 6.0a Time sequence network pattern of cluster (a) I – II with nodes weighted by index

Fig 6.1a Node Probability

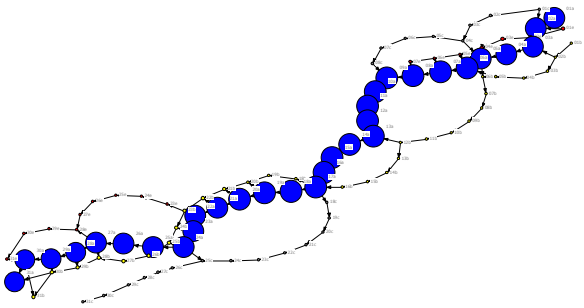


Fig 6.2a General equal probability index

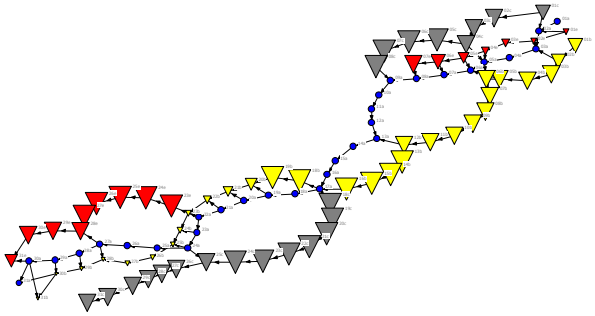


Fig 6.3a General equal probability index A

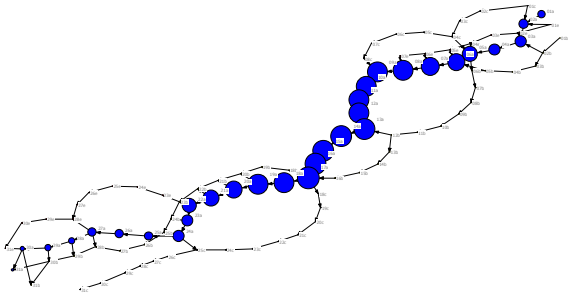


Fig 6.4a Local equal probability index

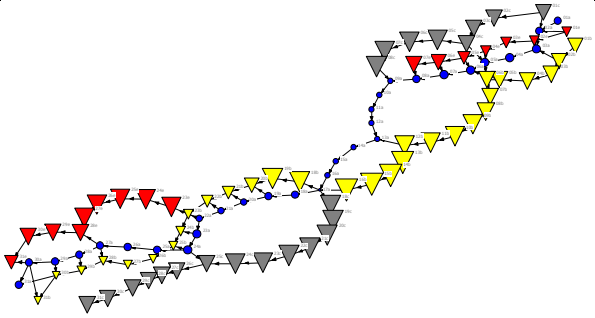
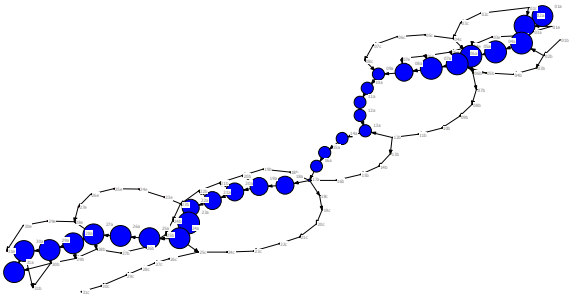


Fig 6.5a Local equal probability index A



The **Local equal probability index** (Fig. 6.4a) is very similar to the above. What changes is that, in this case, the equiprobability is local: it specifies time **t**. At each time **t**, the number of possible events/nodes can change. For example, in the time sequence network of cluster (a) in Figure 5.0, at time $\mathbf{t_1}$ the events/nodes observed, $\mathbf{K_{t=1}}$, are four in number (01a, 01b, 01c, 01e), so that the average number of expected frequencies for the node will be (247/4), while in the sixteenth quarter the events/nodes observed, $\mathbf{K_{t=16}}$, are only two in number (16a, 16b) and the average number of expected frequencies in this case will be (247/2).

$$LEP_{i,t} = ln\left(\frac{f_{i,t}}{\frac{F}{K_t}}\right)$$

As in the previous case (Fig. 6.4a), the shape and size of the node are given by the value and the sign of the index. The graphs in Figures 6.3a and 6.5a show only the nodes with positive index values.

The second group of indexes proposed here measures the density of the ties of each event/node. The first is a classic measure of network analysis and is the **Degree**. This index (Fig. 6.3b) simply counts the number of incoming patterns $\mathbf{k_{i,t-1}}$ and outgoing ones $\mathbf{k_{i,t+1}}$ from each event/node.

$$Degree_i = k_{i,t-1} + k_{i,t+1}$$

The **Local Normalize Degree** index (Fig. 6.1b) measures the ratio between the number of ties that enter and leave node $\mathbf{k_i}$ at time t divided by the number of possible links that can enter and exit node $\mathbf{k_i}$ at time **t**.

$$LNDegree_{i,t} = \frac{k_{i,t-1} + k_{i,t+1} - 2}{K_{t-1} + K_{t+1} - 2}$$

Fig. 6.0b Time sequence network pattern of cluster (a) I – II with nodes weighted by index

Fig. 6.1b Local Normalize Degree

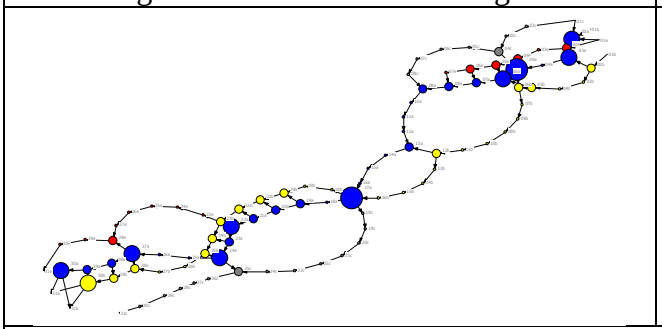
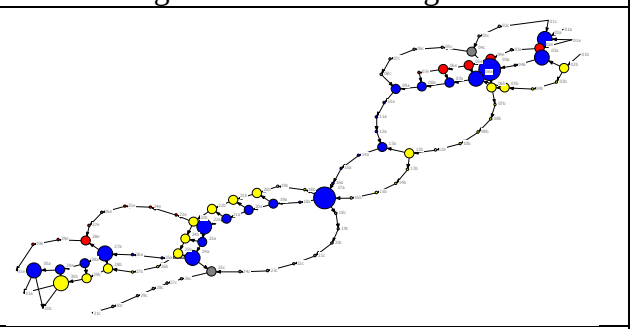


Fig. 6.2b Normalize degree



6.3b Degree

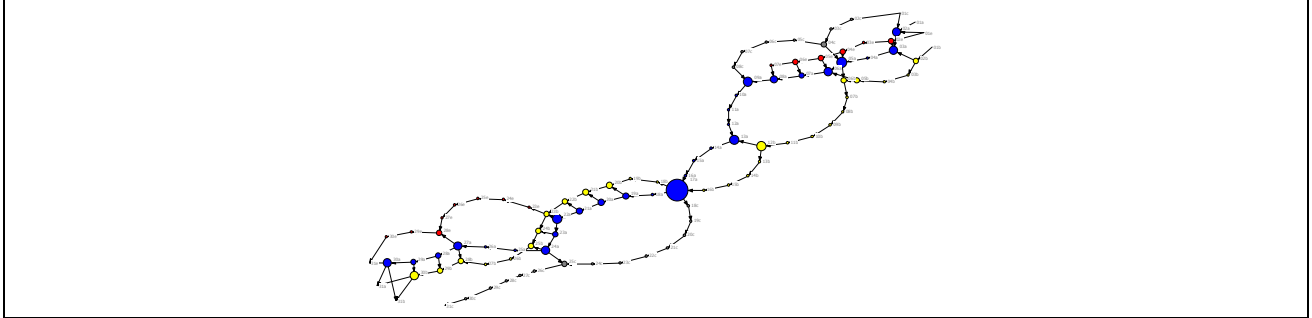


Fig. 6.4b Log degree

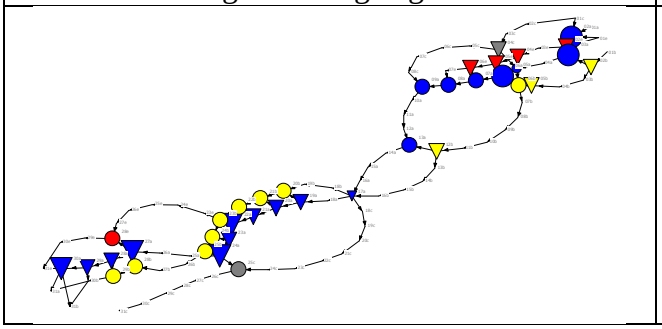
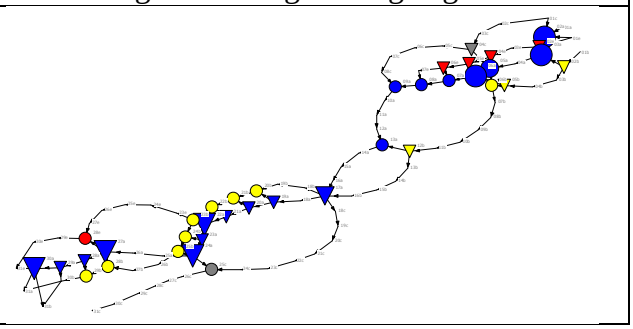


Fig. 6.5b Weighted log degree



The **Normalize Degree** (Fig. 6.2b) is a similar index that normalizes local degree. This is the ratio between the number of ties that enter and leave node $\mathbf{k_i}$ at time **t** divided by two times the total number of possible links/nodes **K**.

$$NDegree_i = \frac{k_{i,t-1} + k_{i,t+1} - 2}{2(K-1)}$$

A different way to observe the links of a node is to compute the ratios between the incoming and outgoing ties: that is, to measure what happens around the event/node $\mathbf{k_i}$ at time **t**. For example, knowing that the number of incoming patterns is greater than outgoing ones may indicate that there has been a simplification of careers at that point in time and a reduction in the system's entropy. The first of the two indices is **LogDegree** (Fig. 6.4b), which is given by the natural logarithm of the ratio between the links in the input and output.

$$LogDegree_{i,t} = ln\left(\frac{k_{i,t-1}}{k_{i,t+1}}\right)$$

The second index is the **Weighted LogDegree** index (Fig. 6.5b), and it is the natural logarithm of the ratio of incoming ties and outgoing ties multiplied by the average number of incoming and outgoing ties.

$$WLD_{i,t} = \left(1 + ln\left(\frac{k_{i,t-1}}{k_{i,t+1}}\right)\right) * \left(\frac{k_{i,t-1} + k_{i,t+1}}{2}\right)$$

Also in this case, used to display the nodes in the network of Figures 6.4b and 6.5b, were the same graphic devices as adopted in Figures 6.2a and 6.2a.

A third group of indexes measure the relationship between all the links between the active time **t-1** and the time **t**, and between time **t** and time **t +1**. In other words, these indices seek to measure the context in which every event/node is inserted.

The first index measures the degree of complexity of the system at various points of observation **t**. The **Local Total Link** (Fig. 6.1c) is the total number of incoming and outgoing ties at time **t**.

$$LTL_t = \sum_{i=1}^{I} k_{i,t-1} + \sum_{i=1}^{I} k_{i,t+1}$$

The second, third and fourth indexes measure the complexity at time **t**. The **Relative indegree-outdegree link** (Fig. 6.2c) is the ratio between all the links in the entry and exit from a node on all possible links existing at that given time t.

$$RIO_{i,t} = \left(\frac{k_{i,t-1} + k_{i,t+1}}{LTL_t}\right) - \left(\frac{1}{K_t}\right)$$

The third is the **Relative indegree link** (Fig. 6.3c), which is the ratio between the ties in the incoming node $\mathbf{k_i}$ at time **t** divided by all ties observed between time **t-1** and **t**.

$$RI_{i,t} = \left(\frac{k_{i,t-1}}{\sum_{i=1}^{I} k_{i,t-1}}\right) - \left(\frac{1}{K_t}\right)$$

The fourth and final index is the **Relative outdegree link** (Fig. 6.4c), and it is the ratio between the ties outgoing from node $\mathbf{k_i}$ at time **t** divided by all active ties between time **t** and time **t +1**.

$$RO_{i,t} = \left(\frac{k_{i,t+1}}{\sum_{i=1}^{I} k_{i,t+1}}\right) - \left(\frac{1}{K_t}\right)$$

Fig. 6.0c Time sequence network pattern of cluster (a) I – II with nodes weighted by index

Fig. 6.1c Total IN OUT	Fig. 6.2c INOUT

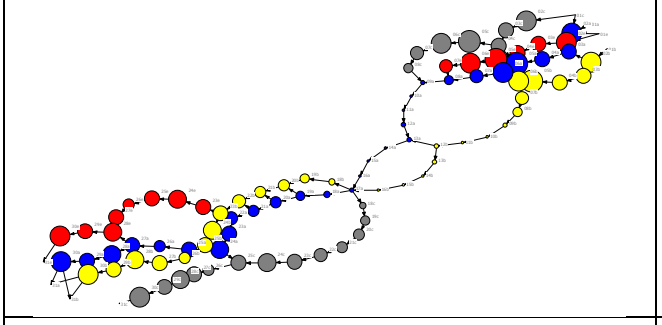
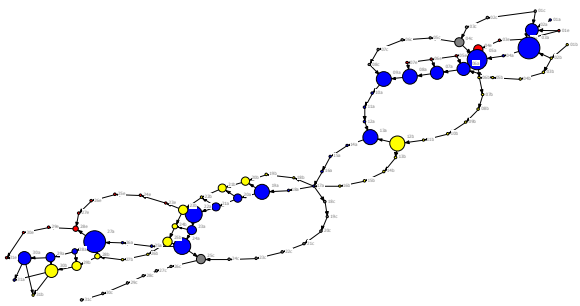
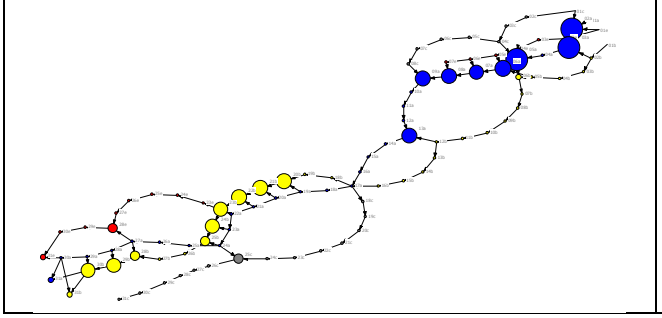
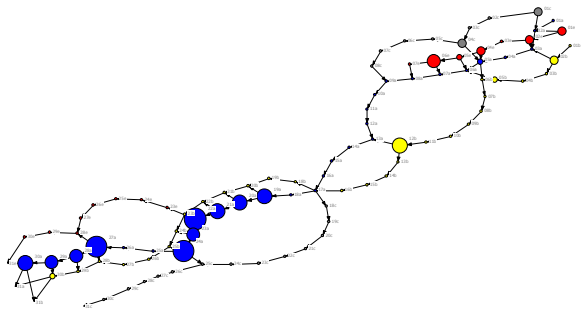

	

Fig. 6.3c IN	Fig. 6.4c OUT
	

The fourth and final group of indices attempts to measure the cutting point. The potentially critical situations that take the form of turning points or critical points in a career: for example, when different patterns collapse into a single pattern/event or, vice versa, when one pattern fragmented in several patterns.

The first index is **Collapse**: (Fig. 6.1d), which is a ratio of ratios that involves the ties incoming to a given node **k** at time **t**. Its value is given by the ratio between the number of incoming ties and the total of the possible incoming ties at time **t** divided by the ratio between all the possible ties at time **t** and the maximum number of possible ties. The value of this index will be higher, the lower the number of ties observed.

$$Col_i = \frac{\left(\frac{k_{i,t-1} - 1}{K_{t-1}}\right)}{\left(\frac{K_{t-1}}{K}\right)} = \left(\frac{K\left(k_{i,t-1} - 1\right)}{2K_{t-1}}\right)$$

The second index is **Sintropy** (Fig. 6.2d). This is the ratio between the total number of nodes and the total number of nodes, at time **t**, minus 1 divided by the total number of nodes minus 1. This index measures the total number of active nodes at a given moment **t** on the total of the possible active nodes. The lower the number of events/nodes observed at time **t**, the greater will be the value of this index.

$$Sint_t = \frac{\left(\frac{K}{K_t}\right) - 1}{K - 1}$$

The last index is **Diversity** (Fig. 6.3d), which is the inverse of the ratio between the total number of links that can be established and the ratio between the number of links in the outgoing of all possible ties that could have been observed between the time **t** and time **t+1**. The index therefore measures the degree of complexity/entropy outgoing from each node/event.

$$Div_i = \left(\frac{K}{\frac{k_{i,t+1}}{K_{t+1}}} \right)^{-1}$$

Fig. 6.0d Time sequence network pattern of cluster (a) I – II with nodes weighted by index

Fig. 6.1d Collapse

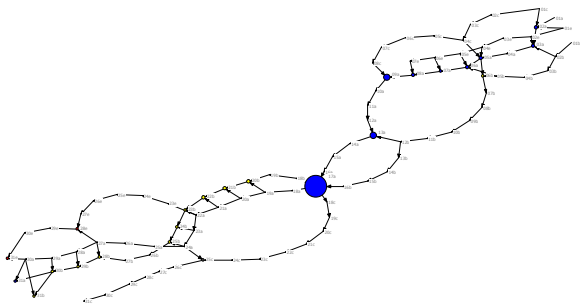


Fig. 6.2d Syntropy

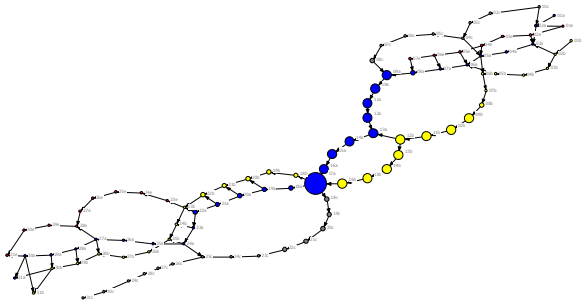
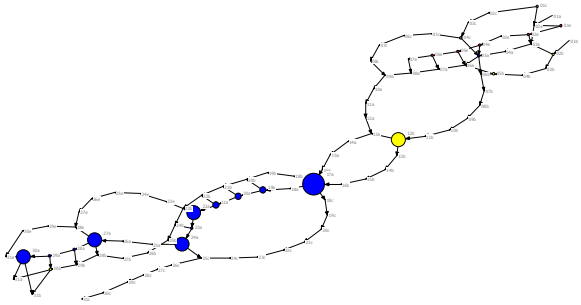


Fig. 6.3d Diversity



As already mentioned, the indices proposed here are only first attempts to measure the sequence-network. Obviously, they should be explored and developed more systematically. For example, some of these measures should be taken into account when calculating the weight of the tie. Account should also be taken of other measures of the nature and contribution of the ties that precede the event, and others which measure the entire structure of the sequence-network. There is still work to be done on these indices.

7.0 Einstein was right. Time is space.

What has been presented is not the only way in which a sequence may be displayed as a network. The newly proposed method is born to the need to preserve the causal structure and the temporal order of events. Each of these graphs is a trace in space of the trajectories followed by a group of actors who move within the time between events. In fact, actors and events do not move through space; they do so through time. They are born, grow up and die over time. Movements between events occur in time, not in space. The transition from one social class to another is not a

physical move from one place to another; rather, it is a change of role (in time). Someway, this graphs describing the motion of the events and actors over time.

The problem is what would happen if we decided to cancel time: in other words, if we decided to eliminate one of the two dimensions of the first network.

To understand what can happen, I shall use as an example the sequences in Table 4.

Table 4.0. List of six random sequences of length seven.

id	t1	t2	t3	t4	t5	t6	t7
5	a	a	b	b	b	b	b
2	a	a	a	a	a	b	b
4	a	a	a	b	b	b	b
6	a	b	b	b	b	b	b
1	a	a	a	a	a	a	b
3	a	a	a	a	b	b	b

Like the previous ones, also these sequences describe the transition between states. They describe in particular the transition between state (a) and state (b). Each is different from the others in at least one element, and this difference can be attributed to a different moment in each of the six sequences; the transition occurs between two states.

Also in this case, the Adjacency Matrix (Table 5.0) will have as many rows and columns as there are distinct elements in the six sequences in question.

Table 5.0 Adjacency matrix of the six random sequences of table 4.0.

	01a	02a	02b	03a	03b	04a	04b	05a	05b	06a	06b	07b
01a	0	5	1	0	0	0	0	0	0	0	0	0
02a	0	0	0	4	1	0	0	0	0	0	0	0
02b	0	0	0	0	1	0	0	0	0	0	0	0
03a	0	0	0	0	0	3	1	0	0	0	0	0
03b	0	0	0	0	0	0	2	0	0	0	0	0
04a	0	0	0	0	0	0	0	2	1	0	0	0
04b	0	0	0	0	0	0	0	0	3	0	0	0
05a	0	0	0	0	0	0	0	0	0	1	1	0
05b	0	0	0	0	0	0	0	0	0	0	4	0
06a	0	0	0	0	0	0	0	0	0	0	0	1
06b	0	0	0	0	0	0	0	0	0	0	0	5
07b	0	0	0	0	0	0	0	0	0	0	0	0

The time sequence network obtained from Table 5.0 is certainly highly distinctive (Fig.7.0). It takes the form of two closely related patterns adjacent to each other. The direction of the transitions is entirely from event (a) to event (b).

What differentiates the two patterns is firstly their different timing. Although both patterns have the same length of time, they are temporally shifted. Pattern (a) occurs an instant before pattern (b), and ends an instant before (b).

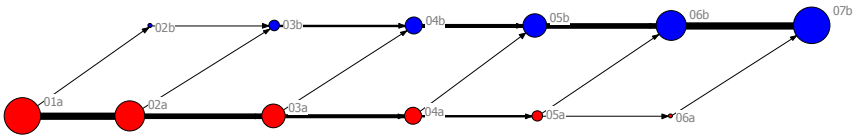
The second difference is that, in the course of time, actors that start in (a), with constant frequency, transit to (b). The result is that the nodes of pattern (b) progressively increase their weight over time, while the nodes of pattern (a) decrease it until its disappearance. Indeed, at the end of the observed process, all subjects initially in event (a) have passed to event (b).

The question at this point is how the graph would change if it was decided to cancel the time dimension. In other words, what would happen if we stopped considering when – at what moment – the transition has taken place and focus only on change of state?

In the meantime, this would significantly reduce the system's complexity. In this new perspective, our attention is directed only to the succession of events as they arise from transitions

between different states within the individual sequences. For example, suppose that the intention is to translate the sequence {aaaabba} into a time sequence network. The distinct nodes of the new graph would be seven in number (a1, a2, a3, a4, b5, b6, a7), as many as the points in time. But if we cancel the time, and then consider only the transitions between different states, the 'new sequence' will be composed of only three nodes (a1, b2, a3), one for each change event.

Fig.7.0. Time sequence network of the six sequences of table 4.0.



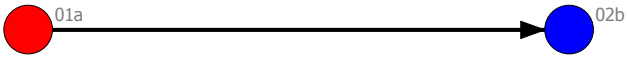
With this passage the one hand we lose the timing of events, but on the other hand we considerably reduce the complexity of the system itself. The point now is to ask what this new object is. What new information emerges from this graph? What factors differentiate it from the time sequence network? What are the limitations and what are the risks of its use?

Table 6.0: Adjacency matrix of the six random sequences in Table 4.0.

	01a	02b
01a	0	6
02b	0	0

Let us return briefly to the sequences in Table 4.0 and build our adjacency matrix (Table 6.0) considering only the changes between different events. In this new definition, this considers the transitions between events, a square matrix with 12 rows and columns changes to one with only two rows and columns.

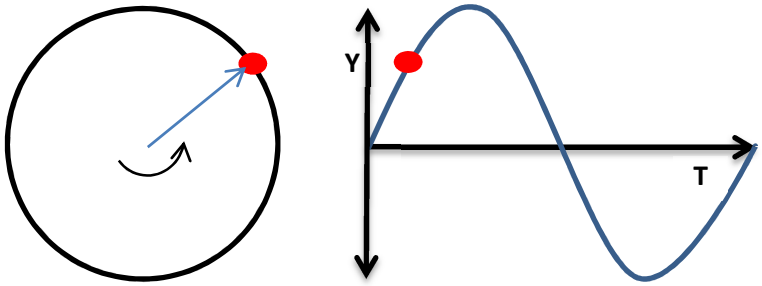
Fig. 8.0: Event sequence network of the six sequences in table 4.0.



Also the network is completely different. This new network (Fig. 8.0), which, in order to distinguish it from the previous one I shall call the event sequence network, consists simply of two nodes and one arc. The graph is quite poor and does not seem to provide much information. Yet this graph is in many ways much more informative than the previous one.

The easiest way to understand what I mean is to recall the first year of high school. In the specific when the physics teacher introduced the simple harmonic motion. Probably he/she would approach the black board and would have drawn a circle. After, he/she went draw a point on the circumference of the circle and the direction of rotation of the point along the circle.

Figure 9.0. Simple harmonic motion:



The teacher, near to the circle, would draw the Cartesian axes Y and T. He/she then would pass to draw on the Cartesian axes the position of the point that travels ideally the circumference. So doing he/she trace a sinusoid or harmonic wave as exemplified in Figure 9.0 which is the spatial representation of the motion of the point in time.

The two graphs are the two sides of the same coin, the same phenomenon. The one is space (motion: the point that travels ideally the circumference) and the other is time (the sinusoid). The former represents the shape of space described at the point in time; the second represents the shape of time described at the point in space.

I think that the relationship that ties our two types of graphs with the example of harmonic motion just presented is now clear. The time sequence network is the sine wave. In space it represents the temporal evolution of the transitions between events. It is the form that changes over time. The event sequence network is the point. It is the shape of space. It is the elementary underlying generative mechanism which produces the sequence over time. They are two sides of the same career. In some ways they are inseparable: the one describes the shape of the space and the other describes the shape of time. Analyze together describes how the space changes over time.

In the case of our last example what are the information that we will obtain. With the time sequence network we have come to the conclusion that the career pattern evolves from (a) to (b) so that it is constant in time. With the event sequence network, we have concluded that the complexity of the entire graph is produced by a single simple generative mechanism which is given by the transition from (a) to (b). Combining the information from both graphs, we can draw the conclusion that pattern (b) is a function of pattern (s).

8.0 Splits Tree: a distance network.

The third type of graph has nothing in common with the previous one. I have decided to include these graphs in this work for four reasons. The first is that this type of graph uses a special representation which is based on the grid network; the second is that this application was created, spread and consolidated within the biomedical disciplines. The third reason is that, among the analysis techniques based on static networks, this is the one which in recent years has undergone the greatest development as regards both the statistical modeling and the graphic parts. And the fourth

one is that this technique graphically displays the distance matrix obtained through the optimal matching or similar procedures.

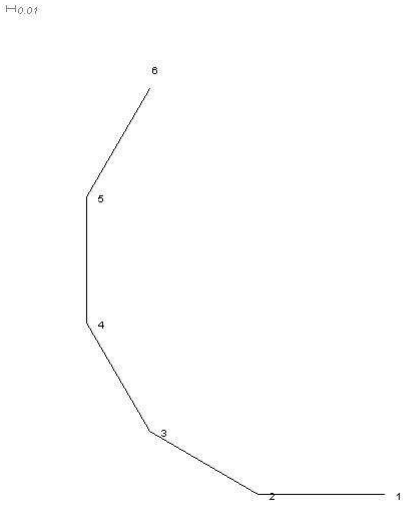
In previous cases, the objects of representation were the individual parts making up the sequence: they were transitions between events or between events in time. These graphical models treat every single sequence as a whole. Secondly, they treat every single sequence as a particular node in a lattice. The relationships between these nodes are determined by their distance, and the link between two sequences is the distance between them.

The simplest distance used in these models is the Levenshtein distance, where the greater the number of elements in common between two sequences, the lower will be the distance that separates the two sequences. However, there are other measures of distance that can be used to model the matrix of the sequences in question.[9]

The result, as said, is a network. At the top are the individual sequences. The lattice describes the structure of the patterns which lead from one sequence to another. The greater the length of the lattice which must be followed to switch from one sequence to another, the greater is the number of elements which do not have the two sequences in common. Examples of the type of configuration that these graphs assume and what information may be found are given below.

In order to show the diversity of the contribution by these graphs to interpretation, and to facilitate their first reading[10], there follow the two networks obtained using the Splits Tree4 program (Huson & Bryant, 2011) which represents the sequences in table 1.0 and 4.0.

Fig. 10. A split network representing the diversity of sequence in Table 4.0



We start from the sequences in Table 4.0 that furnish immediate understanding of the view obtained with these methods. The graph in Figure 10 is a chain, although it has little to do with a chain. In fact, this network connects and sorts sequences according to the number of elements in common. Thus sequences 1 and 2 are close to each other because they have six of the seven elements in common. Also sequences 2 and 3 are close because they too share six of the seven elements composing them. Sequences 1 and 3 are more distant from each other, so much so that we have to cross two segments (1 to 2 and from 2 to 3) to switch between them. This distance is due to the fact that sequences 1 and 3 share only five out of seven elements.

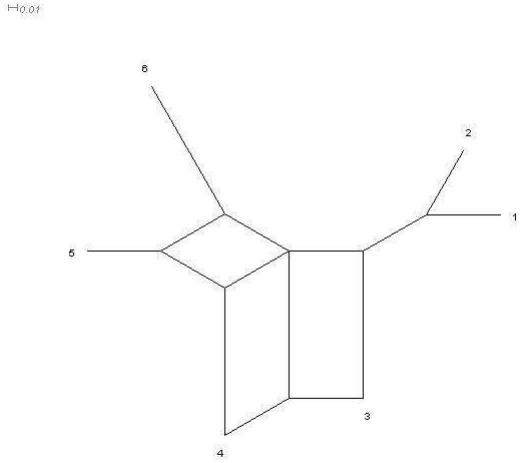
The graph of Table 1.0 is somewhat more complex (Fig.11.0), but its reading does not change. One immediately notices the proximity between the first two sequences and the distance

[9] For a discussion of distances see Huson & Bryant (2011).

[10] For more information on the implementation of these networks and their interpretation see (Wain-Hobson S. et al., 2003; Huson & Bryant 2006).

between the sequences 6 and 1 or 2. As with the previous graph (Fig. 10.0), if we compare each of these three sequences, we find that sequences 1 and 2 have only one different element, while 6 has 3 different elements against 1 and 2.

Fig. 11.0 The split network representing the diversity of sequence in Table 1.0



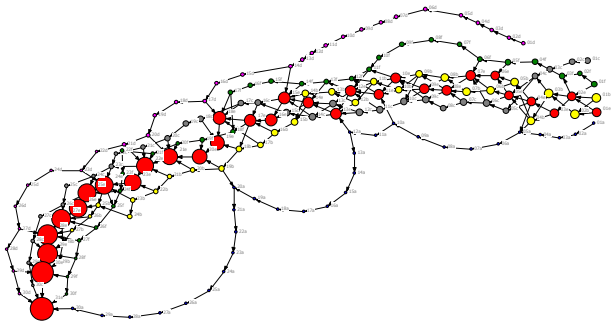
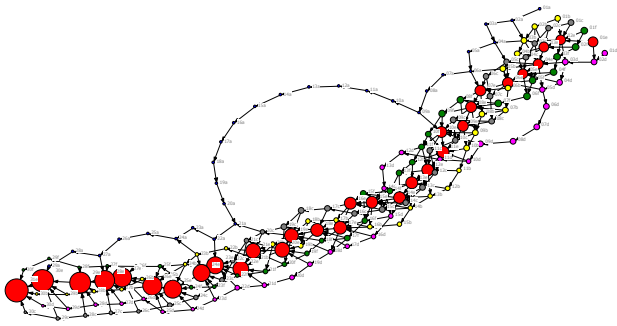
The distinctive feature of these graphs is that they combine the characteristics of a network analysis with the capacity for spatial representation given by multidimensional scaling.

The result is great visual and informative impact where it is not unusual to see very similar structures emerge in clusters, around which are organized the sequences analyzed. Some examples of real data are given below.

9.0 The careers of men and women

I asked in the introduction a series of questions prompted by some remarks on the real capacity of the sequence index plots, and in general display systems, used to capture graphically the structure of the patterns underlying the sequences analyzed.

Fig.12.0 Careers that end in class IIIb+V-VI+VIIIab after ten years.

Fig.12.1 Women (N=126)	Fig.12.2 Men (N=242)
	

I wondered whether the utility of graphical tools could extend beyond simple graphical display; whether a graphical approach could become a valuable investigative tool with which to bring out new ideas on the structure, evolution and composition of the pattern; and whether adoption of these graphical tools could provide a new point of view useful for extending the hypothesis to be tested with other tools.

I tried to answer these questions by proposing the application of network analysis techniques to sequences, and I suggested three distinct types of visualization. Obviously, what follows is only experimental in nature and does not claim to prove anything. It is just another piece of information to test the potential of this approach.

The starting point is the same as in the introduction. I investigate whether the careers of men and women who end in the working class after ten years of work are similar, as could be inferred from observation of the sequence index plot.

Following the order of presentation in the visual analysis, I start with the time sequence network plot (Fig. 12.1 & 12.2).

In contrast with the sequence index plot, visual inspection of the two graphs reveals evident differences between the careers of men and women.

The first finding is the different number of ties in the two networks, as shown in Table 7.0. Women have a smaller average number of links than men. This suggests that women have a career structure simpler than that of men. Women less frequently change from one class position to another and thus have careers which are simpler and less complex than those of men. Having careers structure more simple could also mean having timing and time shape completely different to the men. And that is exactly what is observed when comparing the networks of women and men.

Tab.7.0. Ties and nodes of the career networks of women and man who end in the working class after ten years.

	End in class IIIb+V-VI+VIIa				
	Ties	Nodes	Cases	Mean Ties	Mean Node
Women	283	180	126	1.57	5.81
Men	357	180	242	1.98	5.81
Total	380	181	368	2.10	5.84

The two genders share a common central structure formed by the middle class, the urban petty bourgeoisie, and the urban working class. These three patterns are closely interrelated and share a high number of ties, which indicates that most of the transitions between classes occur between these three classes.

What differentiates them is the timing with which the transitions occur from one class to another. The first difference is the different timings for men and women who start in class I-II. Among women, the first transition from I-II is observed after three years. Thereafter, the few remaining transitions into the other classes occur at intervals of years from each other and do not seem to display some sort of transition pattern. We could say that, among women, descending to the working class after starting in the middle class seems to depend on random factors not related to time.

For men, the structure of the pattern is different. The output from class I-II begins almost immediately and continues with some frequency and regularity in the first three years of work. There are no other transitions from the bourgeoisie I-II to the working class for the next four years. The outflow, however, resumes with regularity in the last three years. Unlike women, apparent in this case is a career pattern with non-random characteristics. The pattern for men is a career in which the risk of descending from class I-II is concentrated at the beginning and end of the observation window. This would suggest that men have a high risk of descending early in their careers where the chances of failure are greater, especially for those men without the means to respond to contingencies that may occur in the earliest years. There is then a period of relative quiet in which the risk of falling diminishes substantially. Finally, before the end of the tenth year of the career, the risk of falling towards the working class starts to increase.

This is not the only difference between male and female careers outlined by the network. Among women, the first transition from IVab is observed after three years; among males it begins

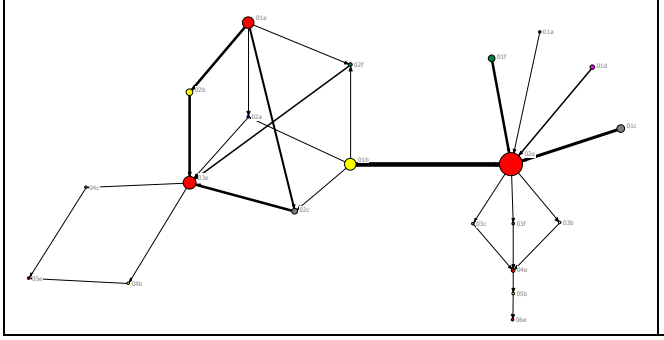
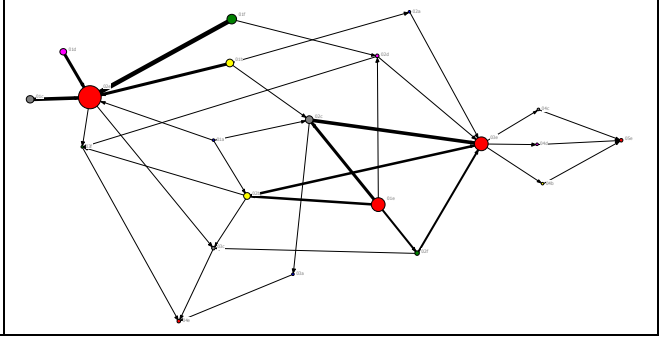
in the first quarter. Also the transitions to and from classes IVab, IIIa, VIIab are significantly more frequent among men than among women. Nevertheless, although the number of ties is smaller, the whole process leading to the working class seems to finish earlier among women. Especially when we consider class IIIa and IVab, we note that most of the transitions from these classes are made until the eighth year of his career. Thereafter, the incidence of these classes is almost marginal in the last two years. Among men, however, there is a continuous flow to and from classes until the last quarter, where by definition all have ended in the working class.

The last feature concerns transitions from the agricultural classes. This pattern is almost non-existent among women. The first transition is observed after about five years, while the other transitions follow the random pattern mentioned above. In contrast, among men there is a dense network of incomings and outgoings from the urban working class. This continuous flow is observed with the same intensity throughout a decade of observation.

Although these observations are not conclusive, they have enabled me to raise a series of issues which if found significant will be tested. However, it is clear that this first application of networks to the study of sequences has already yielded a fair amount of information that, with the instruments used to date, has remained hidden. Consider the possibility of measure the degree of complexity of the system by their numbers of links. With this graphical display we can graphically analyze the evolution of events, capture their timing, and define the time shape in which the patterns unfold.

This, however, is not the only way that we can observe a sequence. A second way is to observe the transitions between events, thereby eliminating the timing of the transitions. Also in this case the difference between men and women is evident.

Fig.13.0 Careers that end in class IIIb+V-VI+VIIIab after ten years.

Fig.13.1 Women (N=126)	Fig.13.2 Men (N=242)
	

Two main patterns are evident in both networks. The first pattern describes the direct transition from classes (a, b, c, d, f) to the urban working class. The second pattern describes those who start in the working class, move into the middle class or the agricultural classes, and then terminate in the working class.

What differentiates men and women is the differing complexity of the network. Women have a simpler network with a lower number of links. Moreover, if we remove the ties from 01b to 02a, 02c, 02f, we obtain two distinct patterns. These results suggest that, among women, there are two primary underlying generative mechanisms which operate separately from each other and combine to form the two main patterns of class careers by women who end up in VIIab after ten years.

The first pattern consists of women who, after entering some class, then move to the urban working class. The second pattern consists of women who start in the working class, pass in the majority of cases to the white-collar middle class IIIb (02b), the urban petty bourgeoisie IVab or into the agricultural working class VIIc and end their journey back in the urban working class.

The male patterns are more complex. In this case it is impossible to identify a single central node that if removed, as in the case of women, produces two separate career patterns. It notes a

significant number of complex patterns with three or more different classes crossed in the first ten years of the career by a large proportion of men.

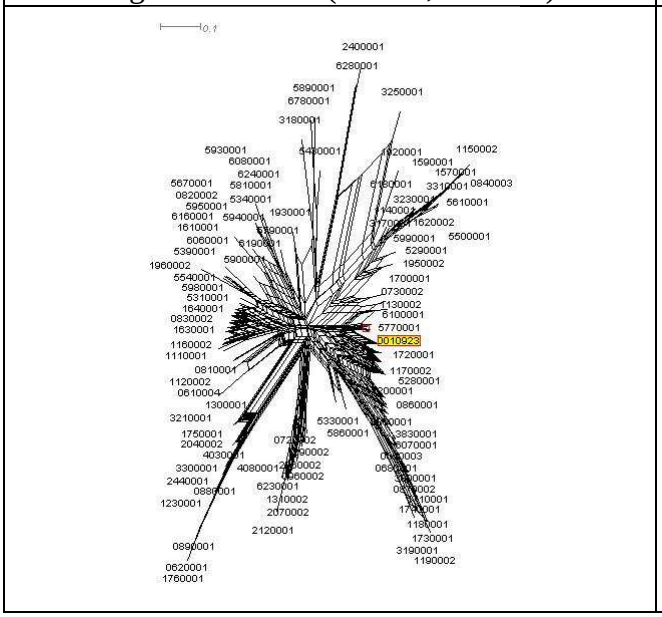
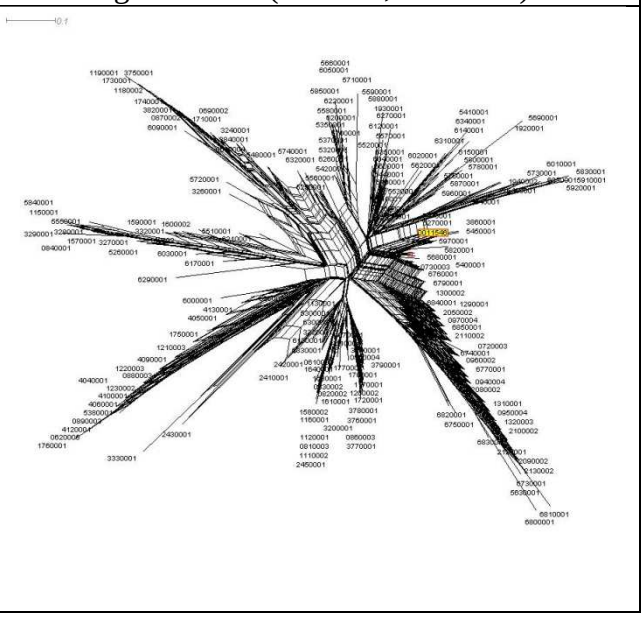
With the third method we again change perspective. Thus far our views have been based on the decomposition of each sequence into its elementary parts. The states and the temporal relations between states have been combined to form networks and events. With this procedure, the sequences have lost their individual identities in favor of a new collective identity made of career patterns. With this third method we resume treatment of the sequences as a whole. The goal is again to use graphical representations to bring out possible common structures underlying the sequences observed.

The main issue to resolve has been how to fit the purpose for which this display system was created with our needs. In the previous sections I repeatedly pointed out that this system was created to display mutations, variations and differences, such as, for example, genetic mutations from a given sequence. The expected input for this type of representation consists of single sequences that differ from each other in at least one item.

The system thus does not handle identical set of sequences. On the other hand, introducing more identical sequences into the analysis would not improve our knowledge about how differences between sequences are articulated. Eventually the information on the number of times a sequence is present could be used to determine what the main diversions are, and then distinguish the main patterns from those that are simply the results of random events.

For this reason, in the example that follows we extract a single copy of each sequence present in the subset of subjects that end in the working class. It is thus clear that the resulting structure does not take account of the different frequencies with which the sequences are present in the sample. All are treated equally, although some of them appear only once and others dozens of times in the same sample.

Fig.14.0 Careers that end in class IIIb+V-VI+VIIIab after ten years.

Fig.14.1 Women (N=923, taxa=99)	Fig.14.2 Men (N=1546, taxa=184)
	

In this case, the graphs resemble stars. At the center are the main sequences, those that have the largest number of elements in common. The more we move away from the centre, the more it becomes clear that the sequences are not distributed uniformly around the centre but along the arms which form the rays of a star. Each ray describes a group of homogeneous sequences because they share a larger number of items among them.

The graph can be conceived as a two-dimensional representation of a multidimensional space of two separate analyses conducted on the same matrix of distances between sequences. The first is show, in this case, by the rays of the star representing groups that can be obtained with a cluster analysis. The second is represented by the distances between points/sequences of the network. Here the distances between centres and their arrangements are the result of a projection obtained by means of a multidimensional scaling of single sequences.

The results of the two graphs are not obtained as clearly and quickly as in previous cases. The number, size and importance of the individual rays of the star seem to be greater for men than for women. This suggests, as also indicated above, that the careers of men are more complex than those of women.

Finally, a distinguishing feature seems to distinguish the males among them. I refer to the group of rays located in the lower quadrant on the left. This group of sequences is almost completely detached from the rest of the star. In other words, there is a group of workers whose career sequences are totally different from those of other men. It is as, if these group of men had a behavior totally different from the others in the mobility processes.

10.0 Conclusions.

This paper has presented the first results from application of network analysis to the study of sequences. It is an introductory work and there are many issues that need to be studied and solved. Nevertheless, a number of insights emerge from these initial experiments that promise well for the future.

First, network analysis seems to be a valuable tool with which to visualize sequences. Through graphs can appear career patterns that are never previously observed. I refer not only to the differing structures of the careers of men and women (already known) used as examples and briefly discussed here. I also refer to the empirical evidence emerging from other analyses of other simple sequences and reported in the appendix.

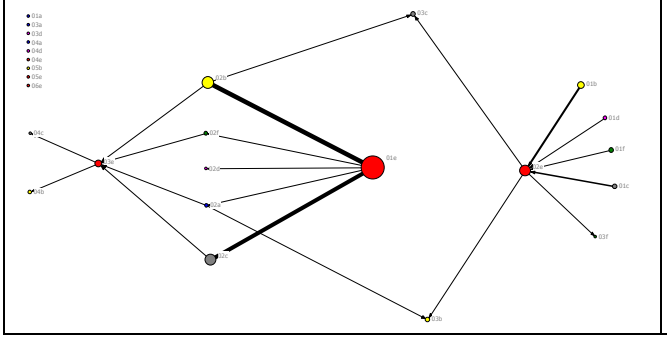
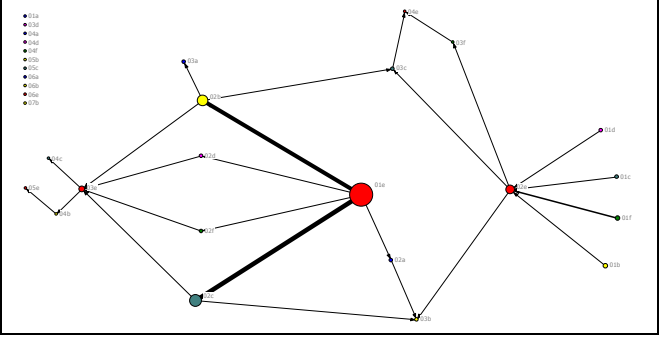
Probably is my mind or my eyes to let me see something that does not exist, but, for example, if we analyze the graph of the cluster (l) (Fig. A.23.0) located in the appendix. The graph at first sight show itself with a fairly complex network structure. Yet, if it is look closely, the network has a clear recursive structure, which starts with one node (b), splits on tree nodes (e, a, c) and back to one node (b). The structure is so clear to allow us to think that the entire graph can be reduced to a simple structure with the form (b, eac, b).

This is not the only situation in which simple recursive structures that act as generative mechanisms underlying sequences are apparent. The examples in the Annex have been chosen at random. The intent was to demonstrate the fallibility of the instrument and its inability to find structures through the network, but this has not happened. Indeed, in many cases, the results were surprising. Take, for example, the graph of cluster (a) of Figure (A.22.0) in appendix. This is a graph composed of a relatively small group of workers (N = 247) yet the structure that emerges is of disarming simplicity.

Clear are also structures that arise when analyzing groups of sequences that share, for example, the beginning or the end of career in the same class. Examples are the patterns of people who began their careers in Class I-II (Figure A.01.0; A.02.0; A.03.0; A.10.0 in appendix). Also in this case, only a few sequences are analyzed, yet the simplicity of the structures that emerge from these trajectories is disarming. Even in the apparently more complex pictures, it is sufficient to exclude those patterns that have a low probability of occurrence in the very clear images.

For example, consider the graphs in the appendix (Figure A.17.0; A.18.0) which describe the sequences of men and women who have spent at least one month in the working class. The number of paths to project is quite large (431 Women, 775 Men) and the two event sequence networks are sufficiently articulated. At first glance it is rather difficult to understand whether the careers of these two groups are the same.

Fig. 15.0 Event sequence network of workers that have spent at least one month in class IIIb+V-VI+VIIIab by women and men.

Women	Men
	

By smaller number of ties would seem that women's careers are simpler. These ties are gathered together around two main trajectories. The first is the pattern of downward mobility exhibited by those women who, after entering some class, descend to the working class. The second is the pattern of upward mobility by those women who leave the working class. In men's careers seem more complex. Clearly apparent is a pattern of upward mobility formed by those men who began their careers in the working class and then moved to the other classes.

However, if we exclude those patterns that are crossed by fewer than four actors, the graphs of men and women change dramatically (Fig. 15.0). What emerges in this particular group, which comprises only those who have spent at least one month in the working class and who have changed occupational class at least once in the ten years, is that there are no substantial differences in career trajectories of class between men and women. The two networks have identical configurations except for some specific links among men and women.

A sequence takes place in time, and the time sequence network has also provided evidence of the extreme sensitivity of the instrument even in visualization of the more marginal patterns. The complexity of interpretation of certain networks is certainly a limit of the method here suggested. Some examples of these complex networks are placed in the appendix. This may induce researchers to be discouraged and abandon this approach to displaying sequences. I think this would be a mistake.

Obviously, the use of these forms of display requires more attention from researchers. They must learn how to separate the patterns that actually exist, even if produced by a small number of subjects, from the ties produced by the noise floor due to errors in data collection.

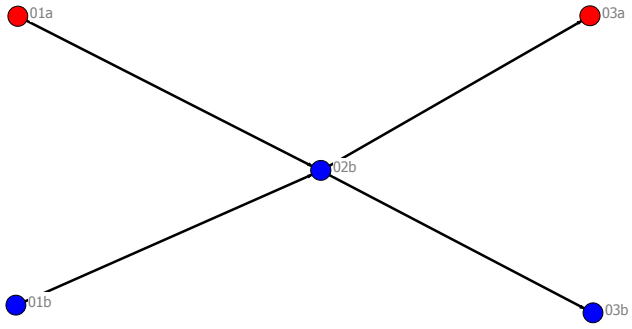
This does not mean that the power of these views is so great as to allow entry into each individual pattern. We can follow the path and its interweaving with other patterns. Through the relationships and transitions from one state to another we can see the timing and time shape of the career pattern in graphical terms.

Obviously, not all graphs are easy to use. Some are so complex and intricate that they resemble more a plate of spaghetti than a network of sequences. I believe however, that this limitation in the ability to read these networks depends only on our ability to find indices, measures and selection thresholds with which to bring out the underlying coherent structures even in these cases.

This is not the only limitation of the method proposed. Among the others, the one that creates the greatest methodological problems is the annulment of individual sequences. This may not be a problem if all actors follow the same pattern, but unfortunately they do not. It may not be a problem if one decides that the past does not affect the future, but this is equivalent to undermining the pillars on which sequence analysis is based. The problem is in these graphs we do not know those who follow who and what. Everything is (con)fused to form a different structure in which the individual trajectories disappear to make space for a ‘mean’ trajectory that describes the transitions between two temporally contiguous points.

Take for example the following sequences: abc, bba, aba, bbc. The resulting network is shown in Figure 16.0, where it is impossible to say who has followed what. What we have here is simply the sum of the various traces left by each actor in covering his/her career.

Fig.16 Time sequence network of sequences: abc, bba, aba, bbc



On the one hand, it can be argued that if a situation like the one just described were real, it would prove to have complete randomness in the sequences and hence a total lack of common patterns. On the other hand, we assume that there are underlying generative mechanisms which produce shared structures (patterns) that will be followed, even if not perfectly, by all the actors analyzed. Each actor may differ by a certain amount but the structure as a whole remains stable. The patterns in these cases constitute the average of individual careers.

What is important in these phases is to not lose sight of individual careers and how they contribute to the realization of the network. A simple method which can be used for this purpose is to use the attributes to define additional information on the nodes. This would (visually) check the disconnect between the performance obtained with the network and the actual one observed in individual sequences

There are many things that remain to be done. Let us consider all the potentialities that here have only briefly mentioned as the biomedical approaches with phylogenetic networks and splits trees method. The appendix reports some examples of this type of analysis conducted on the same set of sequences analyzed above (Figure A.25.0; A.26.0; A.27.0; A.28.0).

This article has illustrated a new, network-based strategy with which to represent and analyze sequences, but this is only one aspect of what can be done by combining networks analysis and sequences analysis. The real novelty, which is not considered in this paper for reasons of space and time, is the transition from static to dynamic.

The real breakthrough will come when we are able to move from a static to a dynamic representation of our phenomena; when our patterns begin to take life and shape before our eyes.

The progress achieved in recent years by network analysis is impressive both in terms of methodology and in technical terms. There is established research areas aimed to modeling networks dynamics. The frontier in this field is no longer that of representing a structure of relations as a whole. The new frontier is representing a structure of relations as a whole which changes over time and space. But this is exactly what we ourselves are trying to do, from another point of view, with the sequence analysis.

Now required are new tools with which to study and graphically model the evolution in time and space of our patterns. I believe, perhaps wrongly, that this opportunity, for now, will be provided by adjusting the tools of network analysis to the study of sequences. I think, in fact, that the shift to a network perspective could provide research tools more useful for studying sequences.

Finally, I think that when we are able to adopt these tools, the only limitation will be our imagination, our ability to imagine.

Reference:

Abbott, A. (1990), *Conceptions of Time and Events in Social Science Methods*, in «Social Science History», 23, 140-50.

Abbott, A. (1992). From causes to events: Notes on narrative positivism. Sociological Methods and Research., 20(4), p.428–455.

Abbott, A. (1995) *Sequence analysis: New methods for old ideas*, Annual Review of Sociology, 21, 93–113.

Abbott A, Forrest J (1986), *Optimal Matching Methods for Historical Sequences*, Journal of Interdisciplinary History, 16, 471-494.

Abbott, A. and Hrycak, A. (1990), *Measuring resemblance in sequence data: An optimal matching analysis of musicians" careers*, American Journal of Sociology, 96(1), 144–185.

Abell, P. (1987). *The Syntax of Social Life: The Theory and Method of Comparative Narratives*. Clarendon Press, New York.

Abell, P. (2004). *Narrative explanation: An alternative to variable centered explanation?* Annual Review of Sociology, 30, p.287–310.

Bateson G.,(1984), *Mente e Natura*, Milano Adelphi. Original title: *Mind and Nature: A Necessary Unity* (1979)

Barkey, K.,Van Rossen R., (1997) *Networks of contention: Villages and regional structure in the seventeenth-century Ottoman Empire*, American Journal of Sociology 102(5). p.1345-1382.

Bearman, P. (1993), *Relations into Rhetorics*, ASA Rose Monograph Series. New Brunswick, NJ: Rutgers University Press.

Bearman P., (2002), *Narrative Networks*, Sunbelt International, Sunbelt Social Network Conference New Orleans, Louisiana February 13-17, 2002, p.89.

Bearman P., Faris R., Moody J., (1999), *Blocking the Future: New Solutions for Old Problems in Historical Social Science*, Social Science History, Vol. 23, No. 4, Special Issue: What Is Social Science History?(Winter, 1999), pp. 501-533

Bearman P., Stovel K., (2001), *Becoming a Nazi: A model for narrative networks*, Poetics 27, 69-90

Bearman P., Moody J., Faris R., (2003), *Networks and History*, C O M P L E X I T Y, Vol. 8, No. 1, p. 61-71.

Berchtold A, Raftery AE (2002). *The Mixture Transition Distribution Model for High-Order Markov Chains and Non-Gaussian Time Series*. Statistical Science, 17(3), 328-356.

Billari FC. (2001), *The Analysis of Early Life Courses: Complex Descriptions of the Transition to Adulthood*, Journal of Population Research, 18(2), 119-124.

Billari FC, Furnkranz J, Prskawetz A (2006). *Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach*. European Journal of Population, 22(1), 37-65.

Bison I., (1999), *Life-packaging in Italy*, paper presented at the POLIS Project conference, Max Planck Institute for Human Development, Berlin, 17-18 March 1999.

Bison I., (2006), *When She Helps Him to the Top*, Conference on Intergenerational transmissions: cultural, economic or social resources? RC28 Spring meeting Nijmegen, 11-14 May, 2006

Bison I., (2011a), *Lexicographic index: A new measurement of resemblance among sequences*, in M. Williams and P. Vogt (eds), The SAGE Handbook of Innovation in Social Research Methods, London, SAGE, pp. 422-441

Bison I., (2011b), *Education, Social Origins and Career (Im)Mobility in Contemporary Italy: A holistic and categorical approach*, in EUROPEAN SOCIETIES, v. Volume 13, n. Issue 3, p. 481-503

Bison I. Esping-Andesen G., (2000), *Life-packaging: Dynamics of working career and family formation in Italy*, Workshop participation of the POLIS project on Globalization, 10-11 March 2000, Madrid, Spain.

Byung-Jun Yoon, Xiaoning Qian, Sayed Mohammad Ebrahim Sahraeian,(2012), *Comparative Analysis of Biological Networks Using Markov Chains and Hidden Markov Models*, IEEE Signal Processing Magazine, Special Issue on Genomic and Proteomic Signal Processing in Biomolecular Pathways, 29(1):22-34. (http://www.ece.tamu.edu/~bjyoon/journal/SPM_2011.pdf)

Borgatti, S.P., (2002), *NetDraw Software for Network Visualization*. Analytic Technologies: Lexington, KY

Bozek K, Thielen A, Sierra S, Kaiser R, Lengauer T., (2009), *V3 Loop Sequence Space Analysis Suggests Different Evolutionary Patterns of CCR5- and CXCR4-Tropic HIV*. PLoS ONE 4(10), p. 1-14.

Brudner, L., White D., (1997) *Class, property, and structural endogamy: Visualizing networked histories*, Theory and Society 26(2/3), p.161-208.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. Stata Journal, 6(4), 435-460.

Chung W, Savell R., Schütt JP., Cybenko G., (2006), *Identifying and Tracking Dynamic Processes in Social Networks*, Proc. Spie Sensors, and Command, Control, Communications, And Intelligence (C3i) Technologies For Homeland Security And Homeland Defense, V.

Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. (2006) Detecting rich-club ordering in complex networks. Nat. Phys. 2, 110‑115.

Dijkstra W, Taris T, (1995). *Measuring the Agreement between Sequences*. Sociological Methods and Research, 24(2), 214-231.

Elzinga, C. H. (2003), *Sequence Similarity: A Non-Aligning Technique*, Sociological Methods and Research 31(4): 3–29.

Elzinga C.H., Liefbroer A.C., (2007), *De-Standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis*, European Journal of Population, 23, 225-250.

Franzosi R., (2004) *From Words to Numbers: Narrative, Data, and Social Science*, Cambridge UK, Cambridge University Press.

Franzosi R, (2010), *Quantitative Narrative Analysis*, Quantitative Applications in the Social Science, n. 162, Thousand Oaks USA, SAGE Publication, Inc.

Franzosi R., Mohr J.W. (1997) *New Directions in Formalization and Historical Analysis, Theory and Society*, Vol. 26(2/3), Special Double Issue on New Directions in Formalization and Historical Analysis (Apr. - Jun., 1997), pp. 133-160

Franzosi R., Bison I, (2010), *Temporal Order: Sequence Analysis*, in Franzosi R, *Quantitative Narrative Analysis*, Quantitative Applications in the Social Science (162), Thousand Oaks USA, SAGE Publication, Inc. pp. 118-123.

Gabadinho A, Ritschard G, Studer M, Muller NS, (2009), *Mining Sequence Data in R with the TraMineR Package: A User's Guide*, Technical report, Department of Econometrics and laboratory of Demography, University of Geneva, Geneva, URL http://mephisto.unige.ch/traminer/.

Gabadinho A, Ritschard G., Muller N.S., Studer M., 2011, *Analyzing and Visualizing State Sequences in R with TraMineR, Journal of Statistical Software*, 40(4), http://www.jstatsoft.org/

Gardy J.L., et al., (2011), *Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak*, The New England Journal of medicine, 364:730-9.

Gauthier JA, Widmer E.D., Bucher P., Notredame C.(2010), Multichannel Sequence Analysis Applied to Social Science Data, Sociological Methodology,

Gould, R.V. (1995), *Insurgent Identities: Class, Community, and Protest in Paris from 1848 to the Commune*, Chicago, University of Chicago Press.

Gould, R.V., (1996), *Patron-client ties, state centralization, and the Whiskey Rebellion*, American Journal of Sociology 102(2), p.400-429.

Guimera, R., Sales-Pardo, M. & Amaral, L. A. N. (2007) Classes of complex networks defined by role-to-role connectivity profiles. Nat. Phys. 3, 63–69.

Hanneman, R. A., Riddle M., (2005), *Introduction to social network methods*, Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/)

Hanneman, R. A., Riddle M., (2005), *Working with Netdraw to visualize graphs*, Riverside, CA: University of California, Riverside (published in digital form at http://faculty.ucr.edu/~hanneman/nettext/C4_netdraw.html

Herrnstadt C., Elson J.L., Fahy E., Preston G., Turnbull D.M., Anderson C., Ghosh S.S., Olefsky J.M., Beal M.F., Davis R.E., Howell N.,(2002), *Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups*, The American Society of Human Genetics, 70:1152–1171

Huson D.H., Bryant D., (2006), *Application of Phylogenetic Networks in Evolutionary Studies*, Molecular Biology and Evolution, 23(2):254–267.

Huson D.H., Bryant D., (2011), *User Manual for SplitsTree4 V4.12.3*, http://ab.inf.uni-tuebingen.de/data/software/splitstree4/download/manual.pdf

Kalari KR, Rossell D, Necela BM, Asmann YW, Nair A, Baheti S, Kachergus JM, Younkin CS, Baker T, Carr JM, Tang X, Walsh MP, Chai H-S, Sun Z, Hart SN, Leontovich AA, Hossain A, Kocher J-P, Perez EA, Reisman DN, Fields AP and Thompson EA, (2012), *Deep sequence analysis of non-small cell lung cancer: integrated analysis of gene expression, alternative splicing, and single nucleotide variations in lung adenocarcinomas with and without oncogenic KRAS mutations*. Frontiers in Oncology | Cancer Genetics. 2:12.

Kuchaiev O., Milenkovic T., Memisevic V., Hayes W, Przulj N., (2010), *Topological network alignment uncovers biological function and phylogeny*, J. R. Soc. Interface published online 17 March 2010, http://rsif.royalsocietypublishing.org/content/early/2010/03/24/rsif.2010.0063.full.html#related-url

Martin, P., Schoon, I. and Ross, A. (2008), *Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research*, International Journal of Social Research Methodology, 11(3), 179-199.

ONA survey (2009), *NETDRAW – BASIC A Practical Guide to Visualising Social Networks* , Version 1.0, http://www2.optimice.com.au/documents/ONANetdrawGuideBasic.pdf

Padgett, J., Ansell C., (1993), Robust action and the rise of the Medici, 1400-1434, American Journal of Sociology 98(6), p.1259-1319.

Pentland B. T., Feldman M. S., (2007), Narrative Networks: Patterns of Technology and *Organization, OrganizationScience,* Vol. 18(5), pp. 781–795.

Ritschard G, Gabadinho A, Muller NS, Studer M (2008). *Mining Event Histories: A Social Science Perspective.* International Journal of Data Mining, Modelling and Management, 1(1), 68-90.

Rosenthal, N., Fingrutd M., Ethier M., Karant R., McDonald D., (1985), *Social movements and network analysis: A case study of nineteenth-century women's reform in New York State*. American Journal of Sociology, 90 (5): 1022-54.

Scherer S., (2001), *Early Career Patterns: A Comparison of Great Britain and West Germany*. European Sociological Review, 17(2), 119-144.

Srinivasan B.S., Shah N.H., Flannick J. A., Abeliuk E., Novak A.F., Batzoglou S., (2007), *Current progress in network research: toward reference networks for key model organisms*, Briefings in Bioinformatics. Vol. 8. No 5. 318 -332

Stark D., Vedres B., (2002), *Pathways of Property Transformation: Enterprise Network Careers in Hungary, 1988-2000,* Sunbelt International Sunbelt Social Network Conference New Orleans, Louisiana February 13-17, 2002 ,p.74)

Wain-Hobson S., Renoux-Elbe C., Vartanian J-P., Meyerhans A., (2003), *Network analysis of human and simian immunodeficiency virus sequence sets reveals massive recombination resulting in shorter pathways*, Journal of General Virology, 84, 885–895

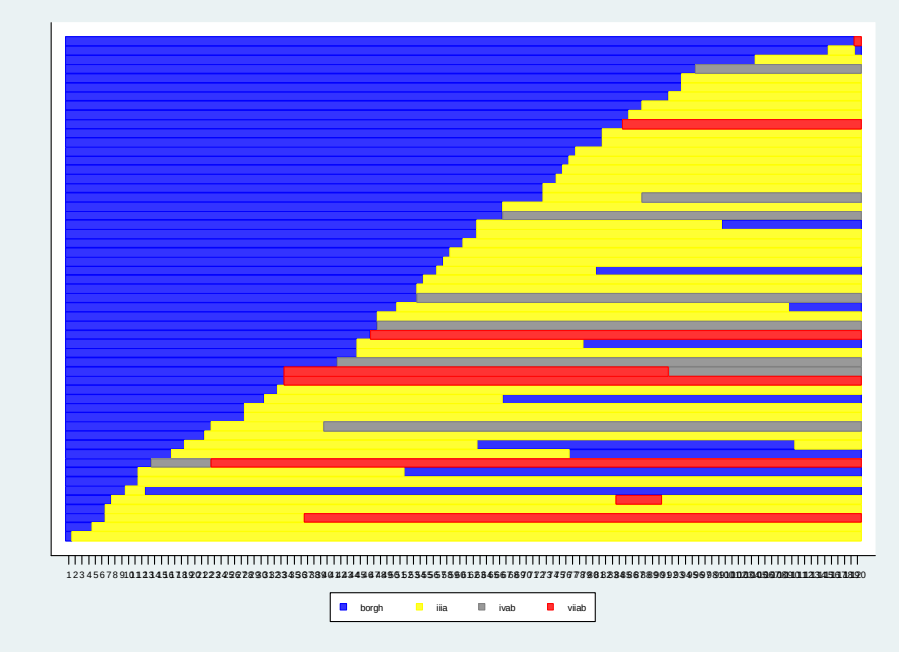
Widmer E, Ritschard G (2009). *The De-Standardization of the Life Course: Are Men and Women Equal?*, Advances in Life Course Research, 14(1-2), 28-39.

Wiggins, R.D., Erzberger, C., Hyde, M., Higgs, P., and Blane, D. (2007) *Optimal matching analysis using ideal types to describe the lifecourse: an illustration of how histories of work, partnerships and housing relate to quality of life in early old age*, International Journal of Social Research Methodology, 10(4), 259-278.

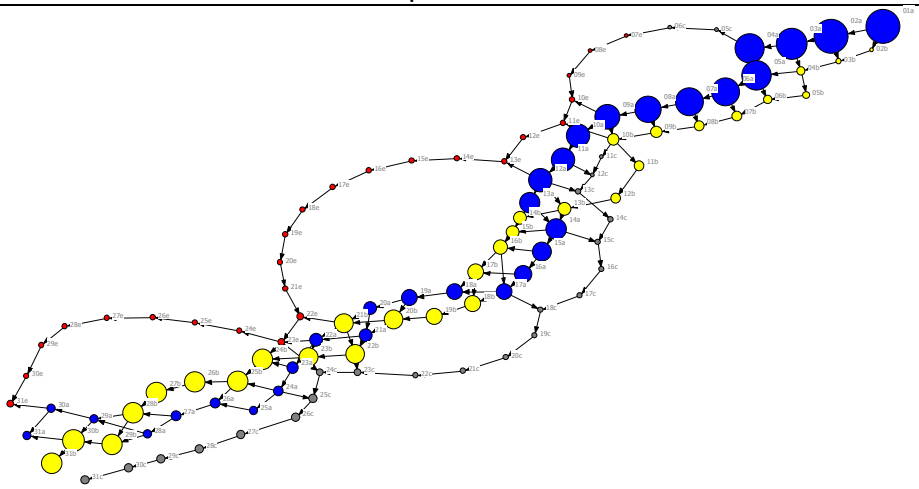
Appendix

Figure A.01.0. Class careers that begin in class I & II. Total (N=55)

Sequence Index plot



Time Sequence Network



Event Sequence Network

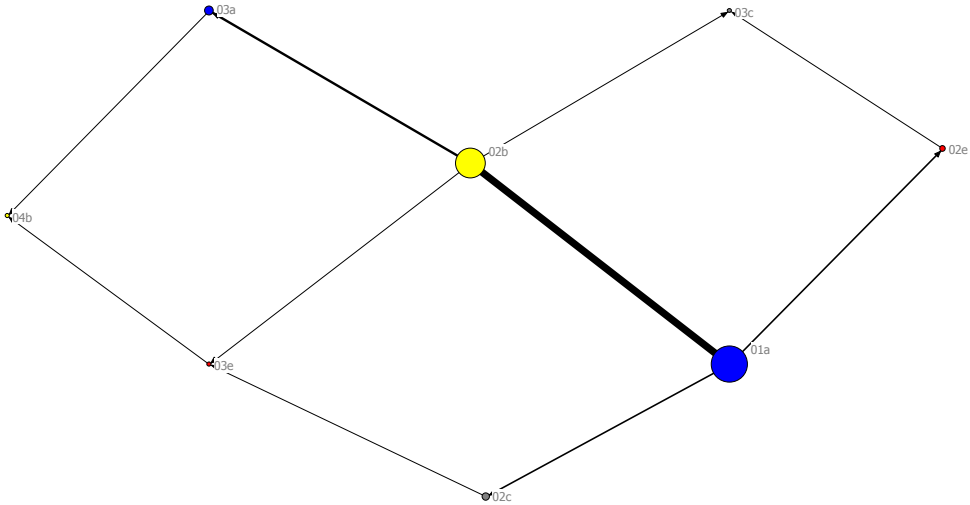
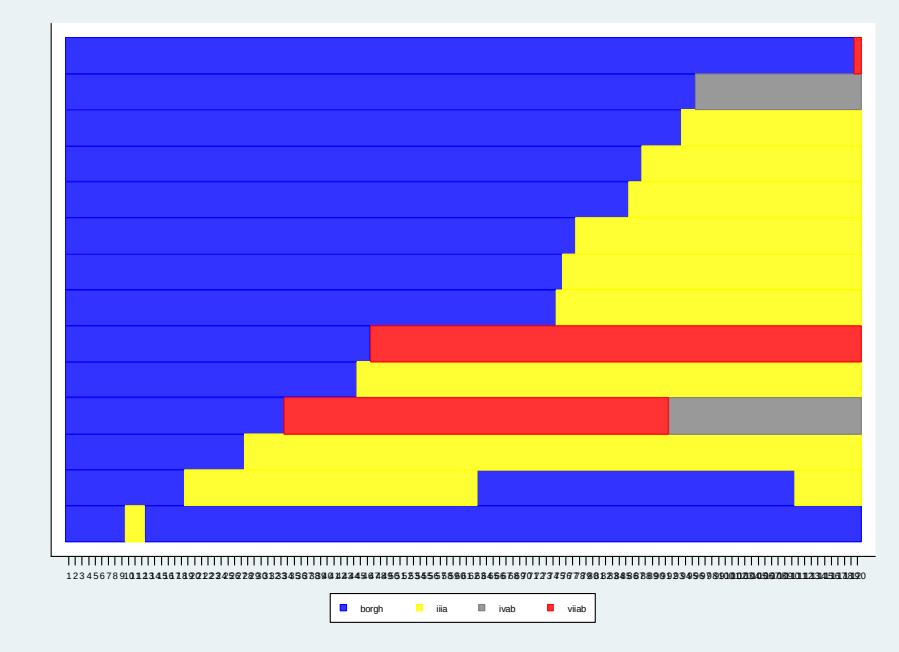
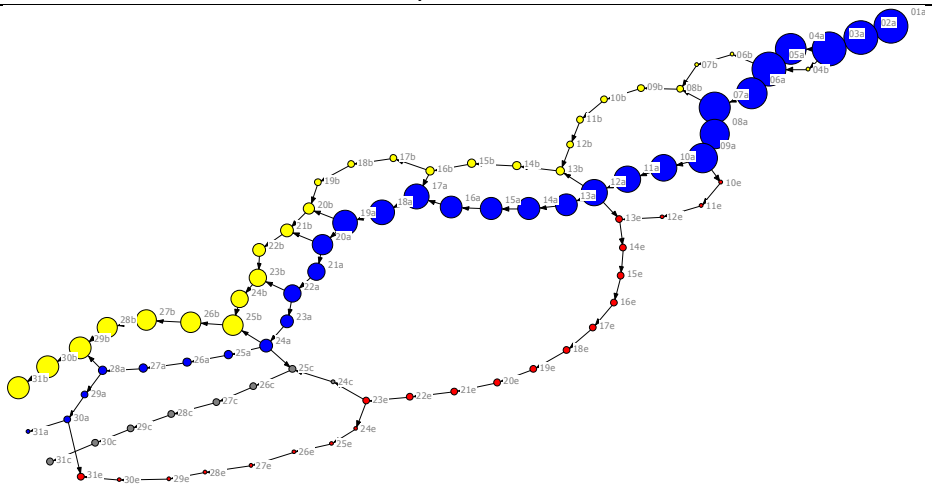


Figure A.02.0. Class careers that begin in class I & II. Women. (N=14)

Sequence Index plot



Time Sequence Network



Event Sequence Network

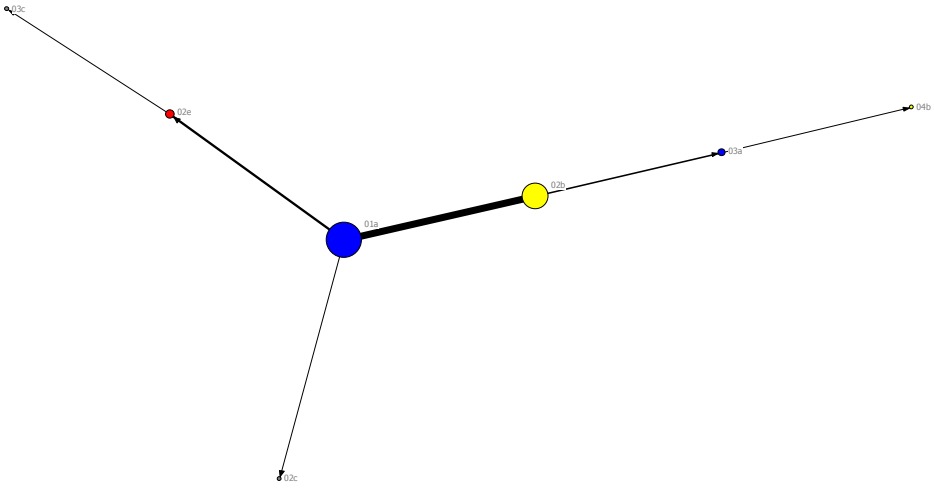
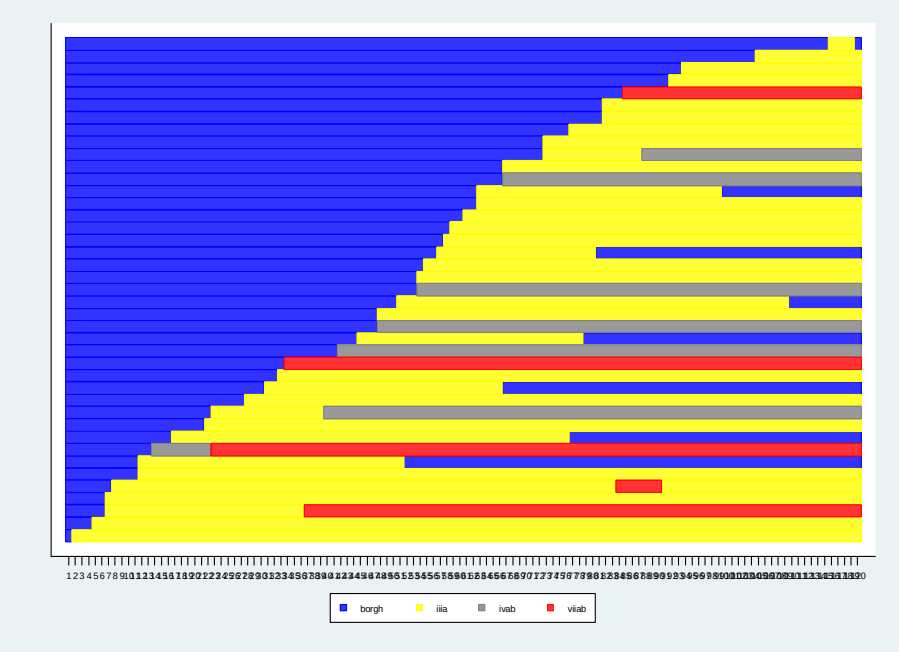
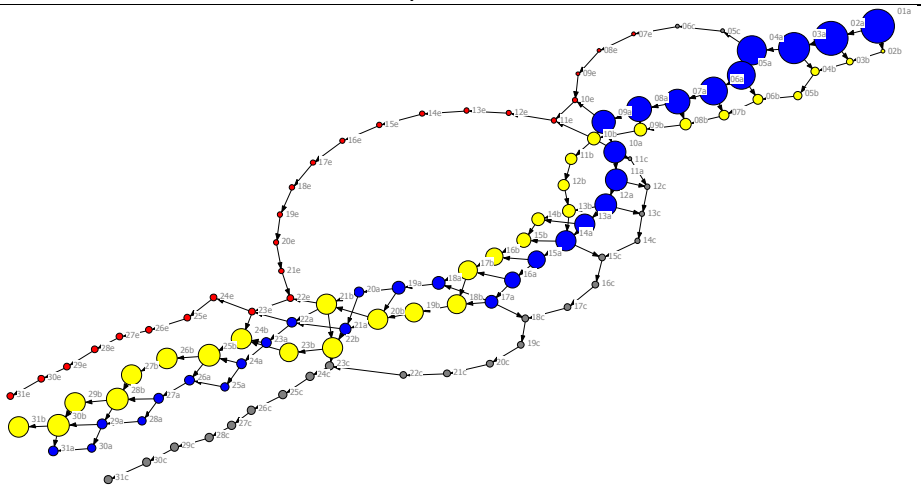


Figure A.03.0. Class careers that begin in class I & II. Men (N=41)

Sequence Index plot



Time Sequence Network



Event Sequence Network

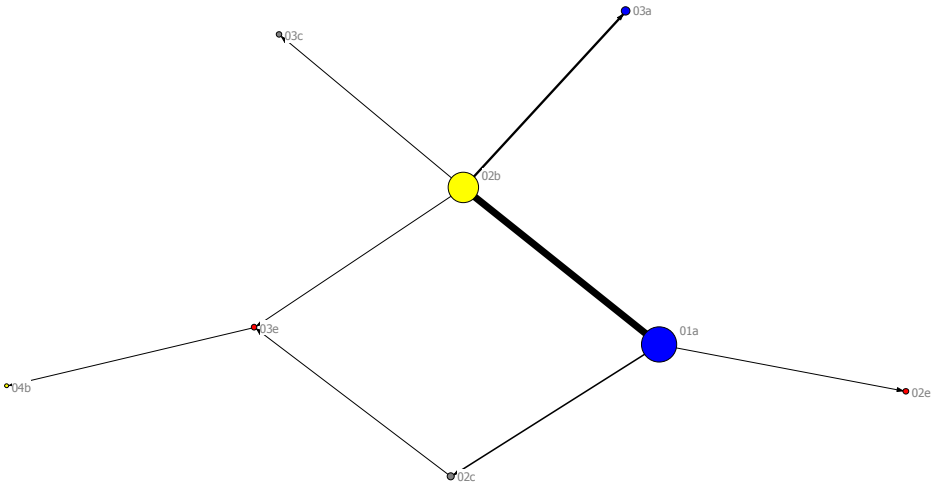


Figure A.04.0. Time sequence network of class careers that begin in class I & II on total and by gender.

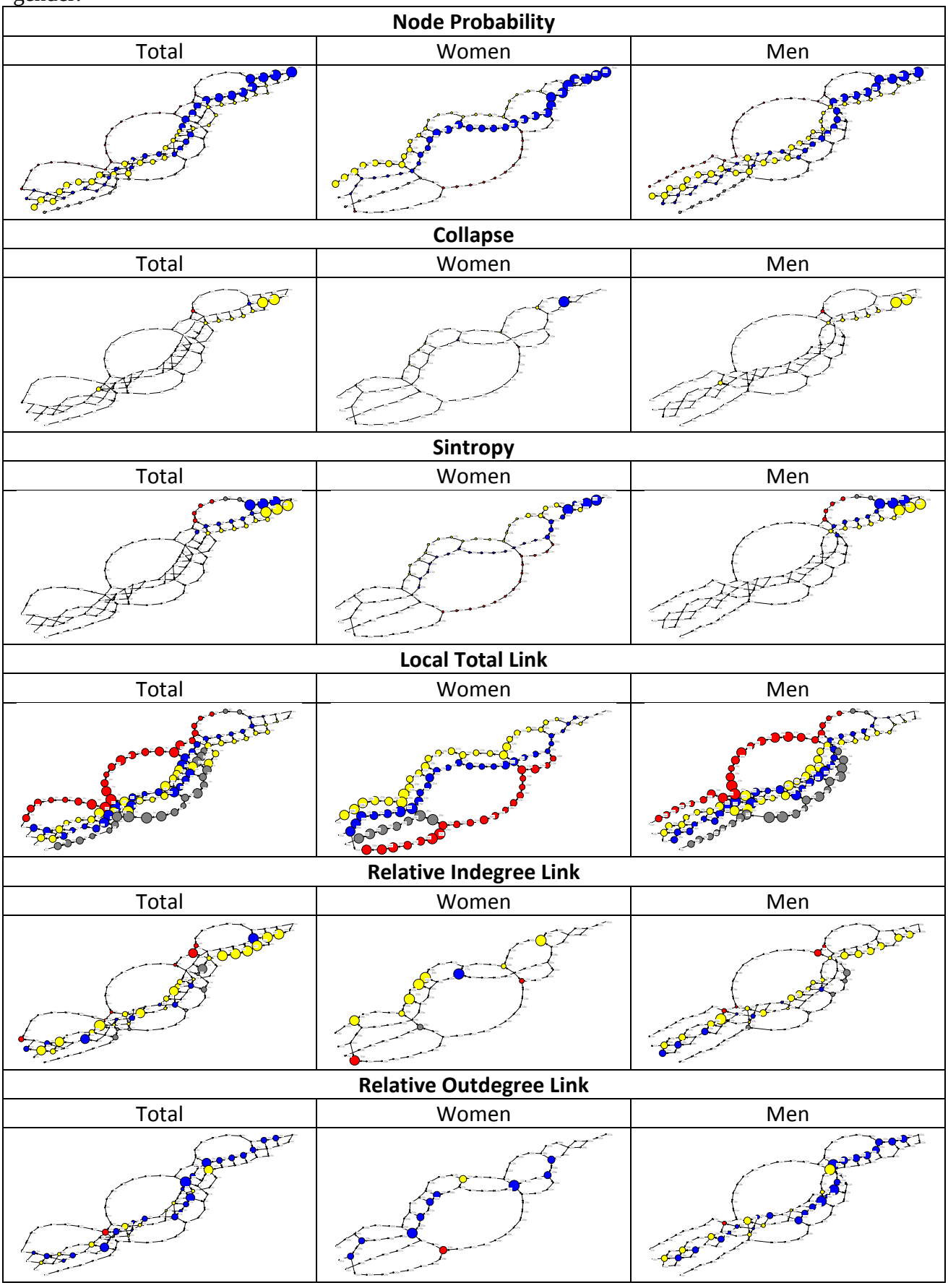
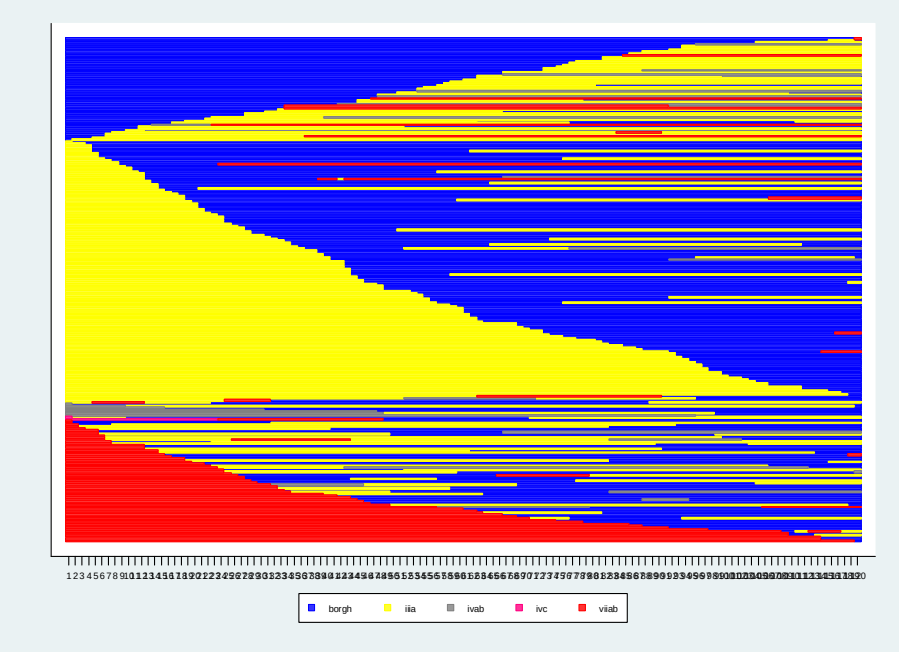
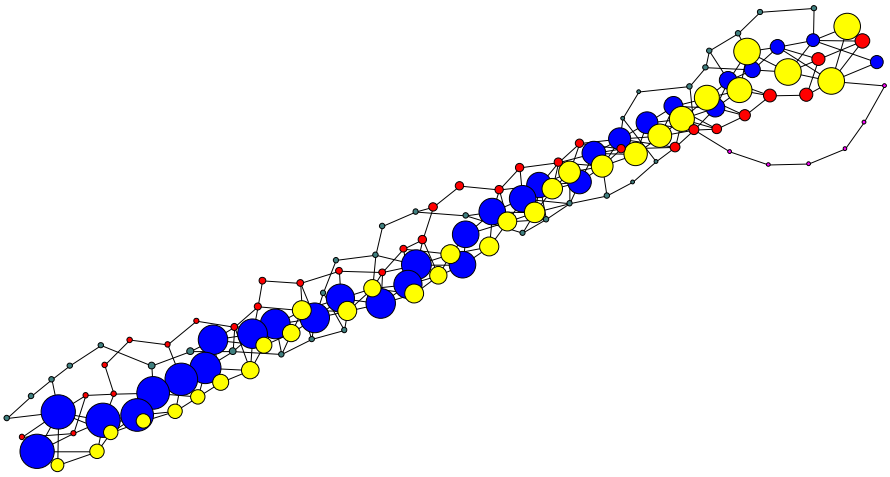


Figure A.05.0. Working class careers where at least one month has been spent in class I & II. Total (N=269)

Sequence Index plot



Time Sequence Network



Event Sequence Network

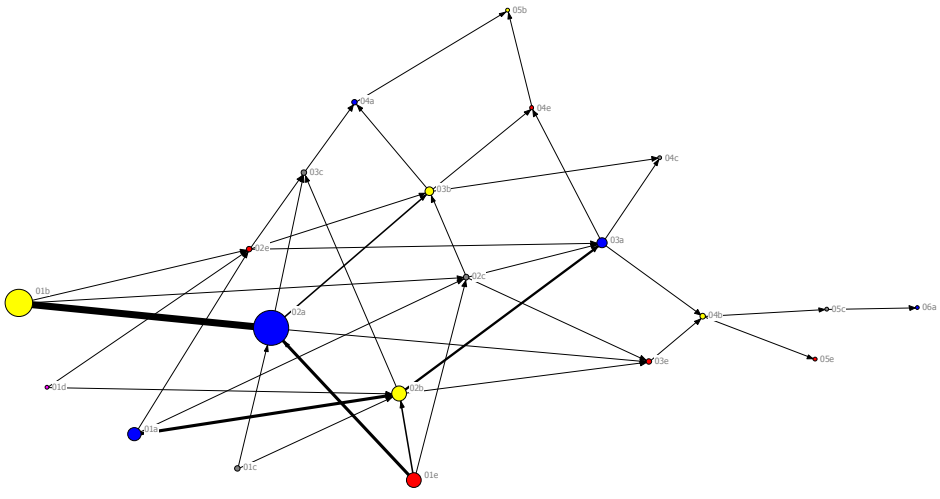
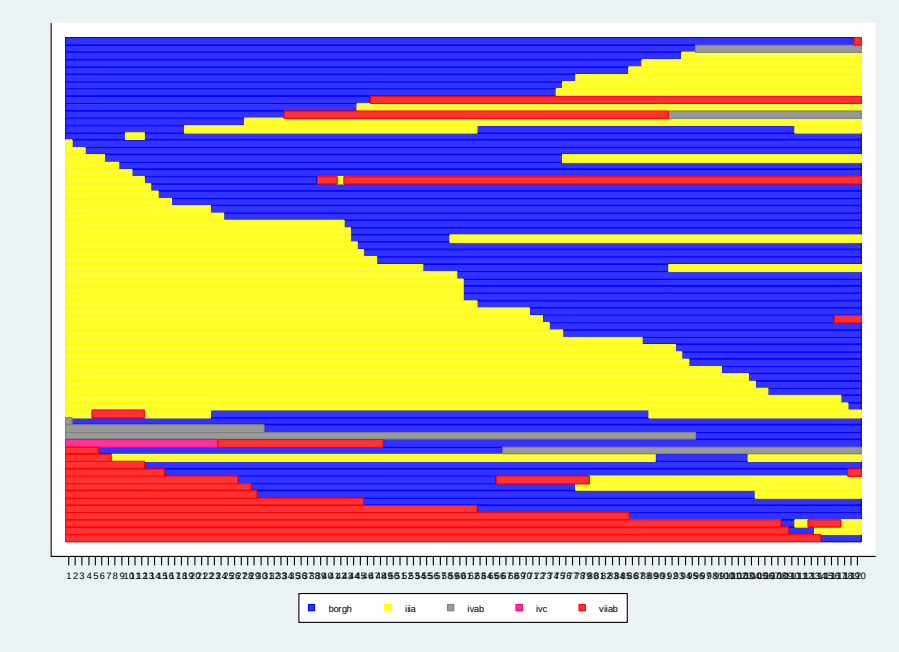
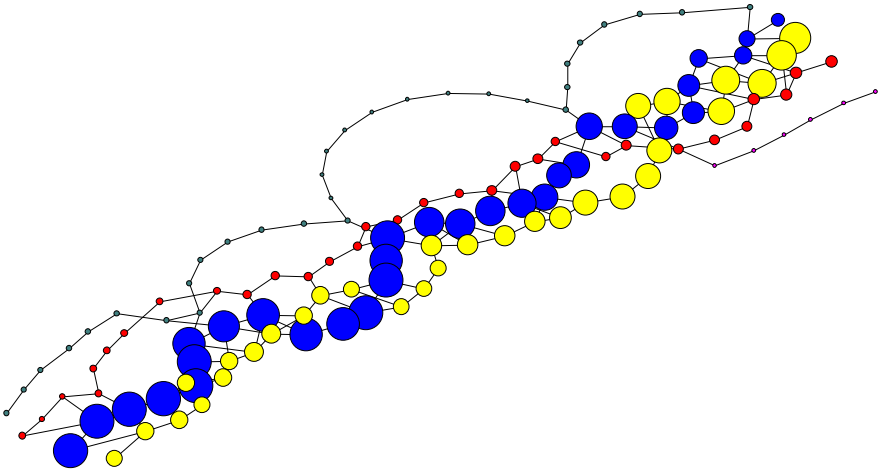


Figure A.06.0. Working class careers where at least one month has been spent in class I & II. Women (N=69)

Sequence Index plot



Time Sequence Network



Event Sequence Network

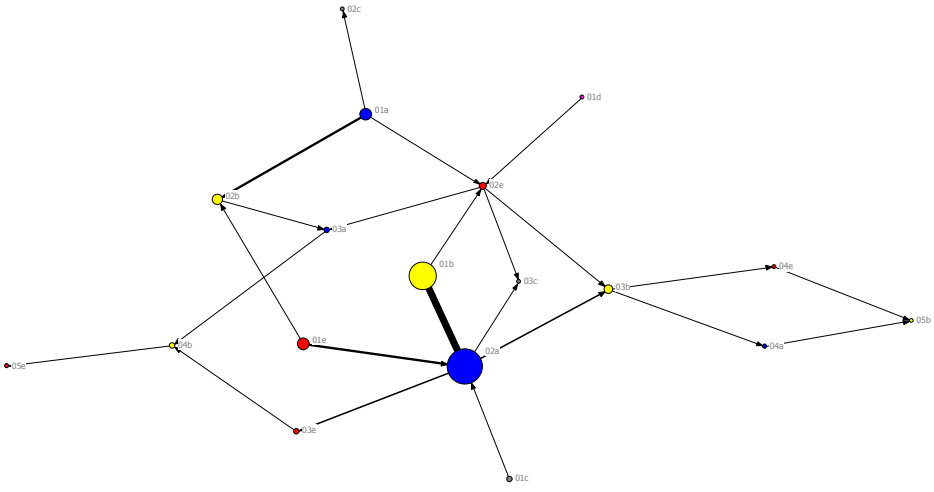
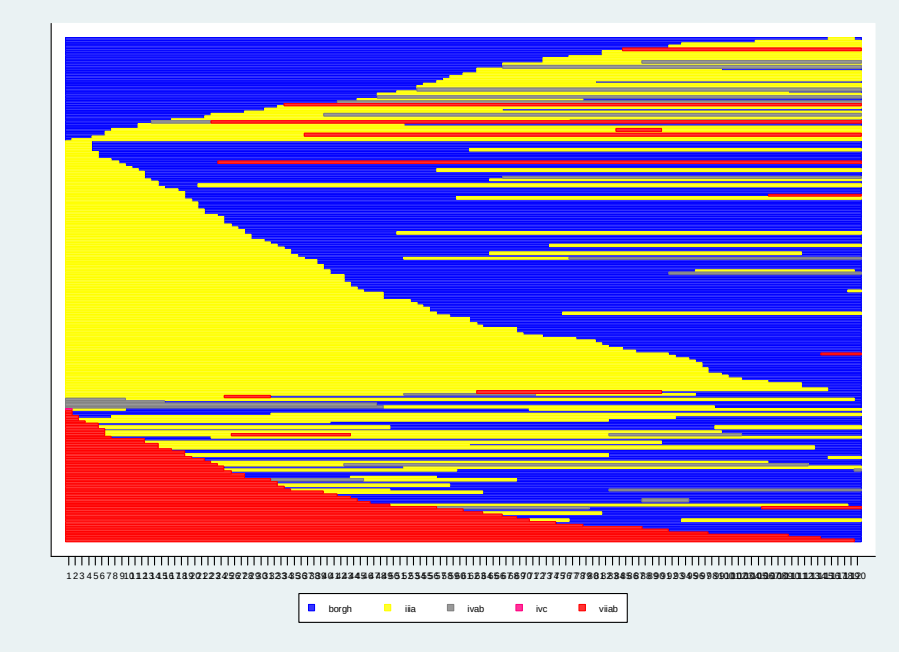
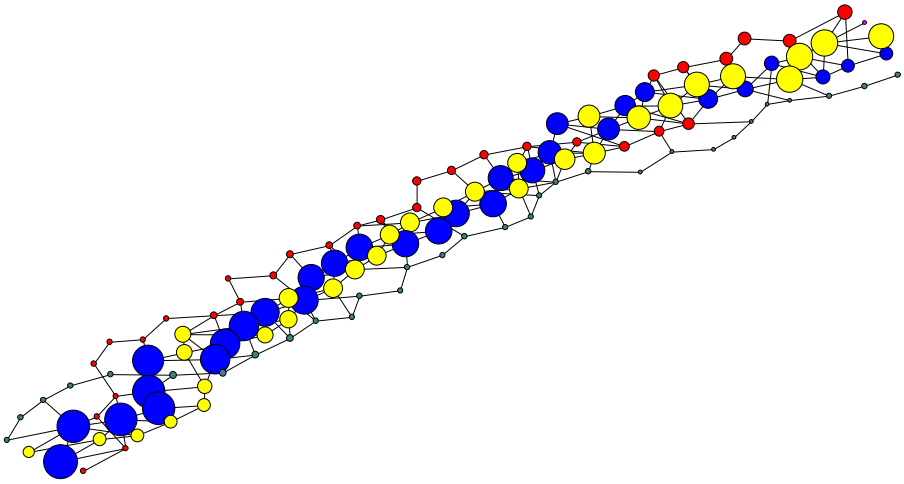


Figure A.07.0. Working class careers where at least one month has been spent in class I & II. Men (N=200)

Sequence Index plot



Time Sequence Network



Event Sequence Network

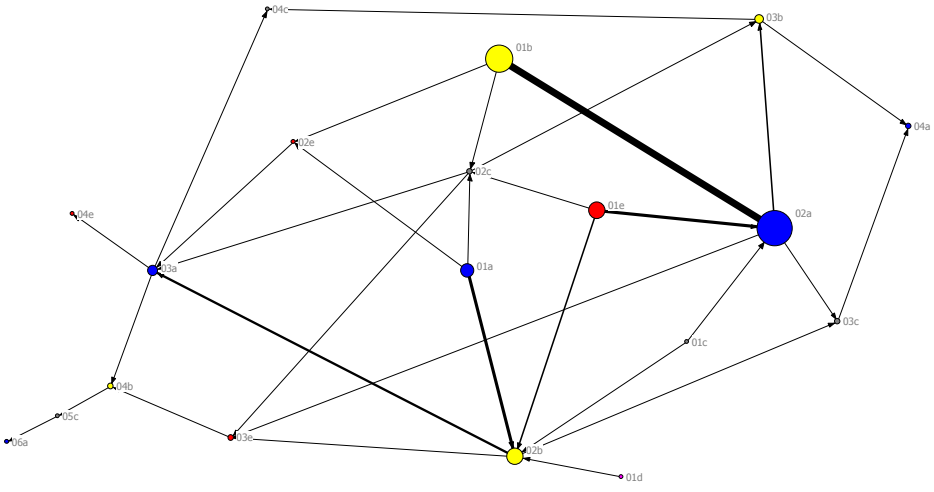


Figure A.08.0. Time sequence network of working class careers where at least one month has been spent in class I & II by gender and some time network measure.

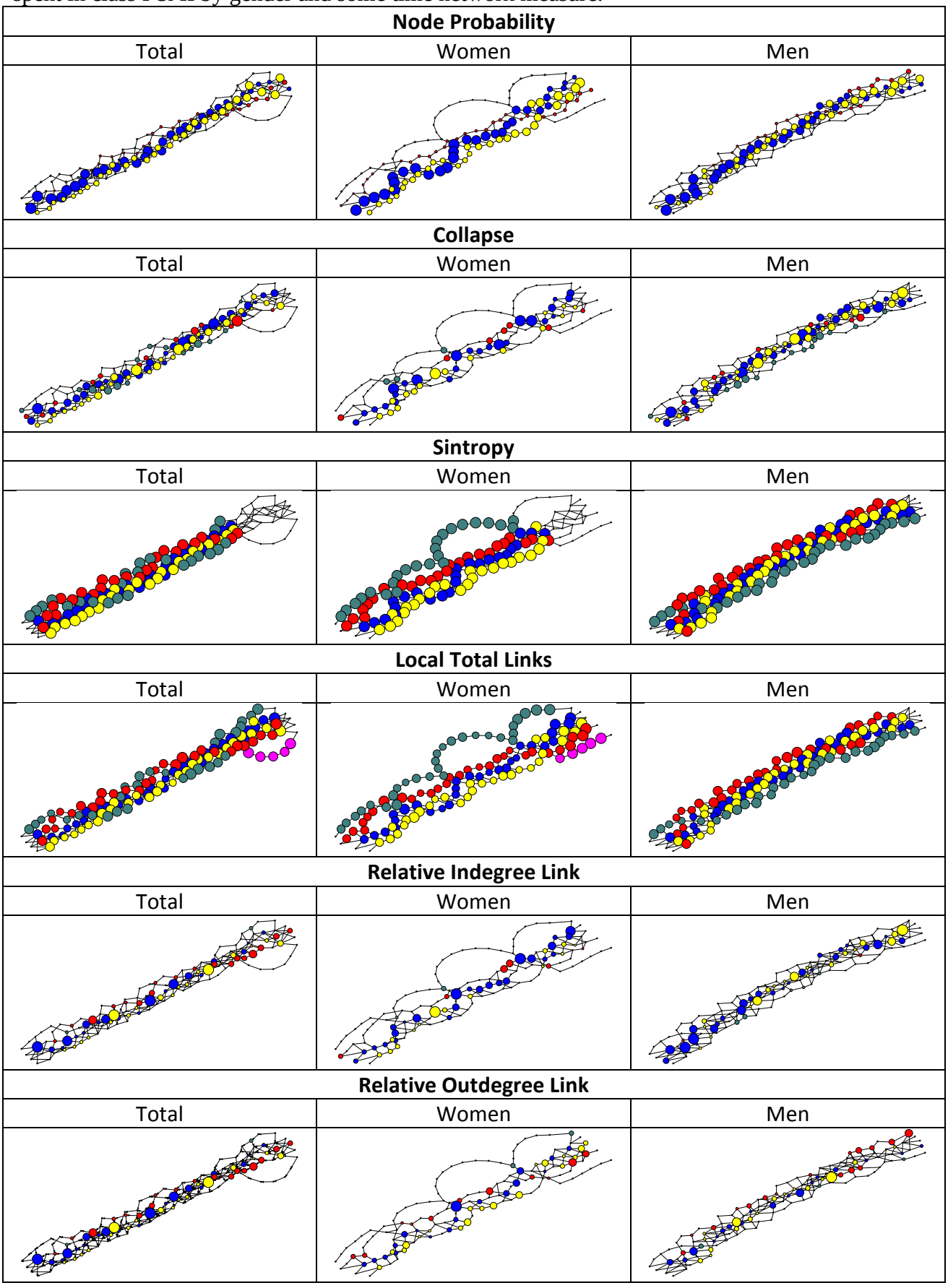
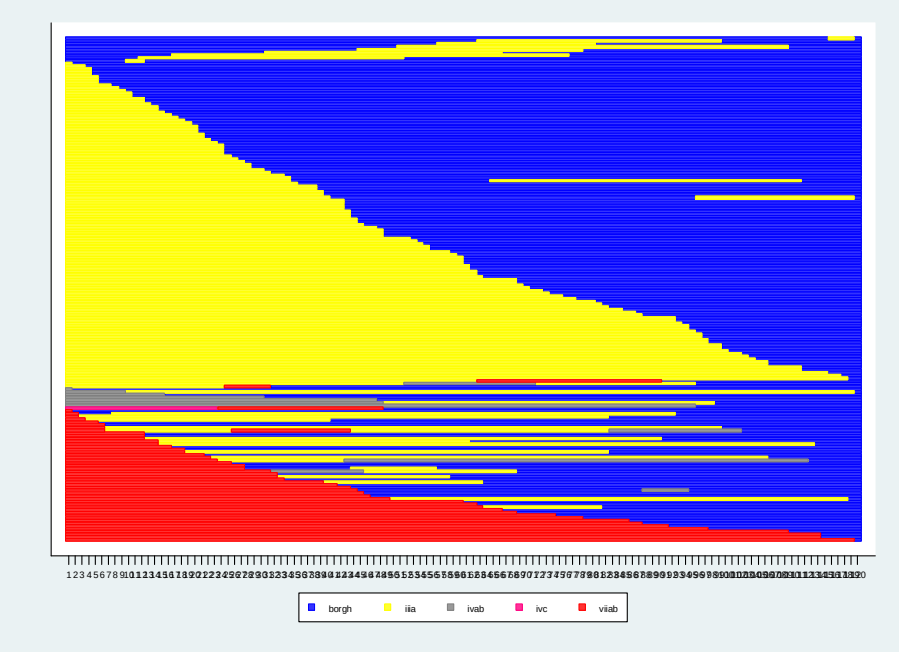
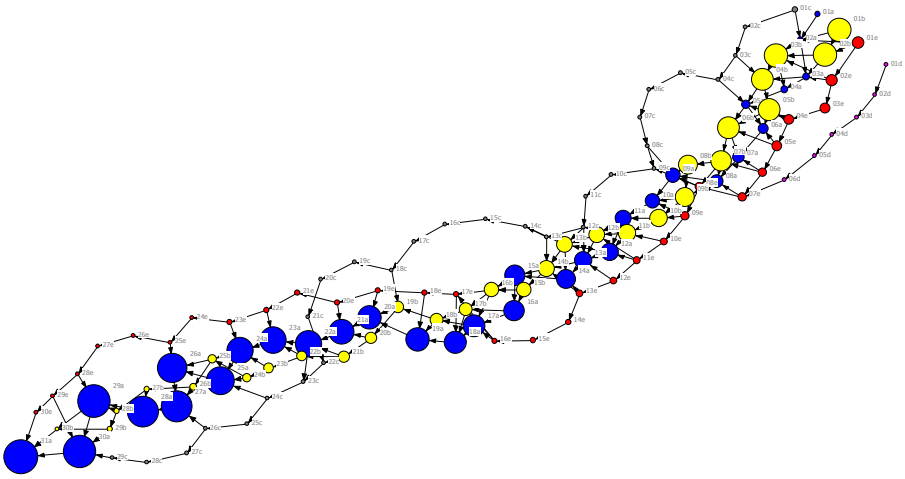


Figure A.09.0. Class careers that end in class I & II. Total. (N=184)

Sequence Index plot



Time Sequence Network



Event Sequence Network

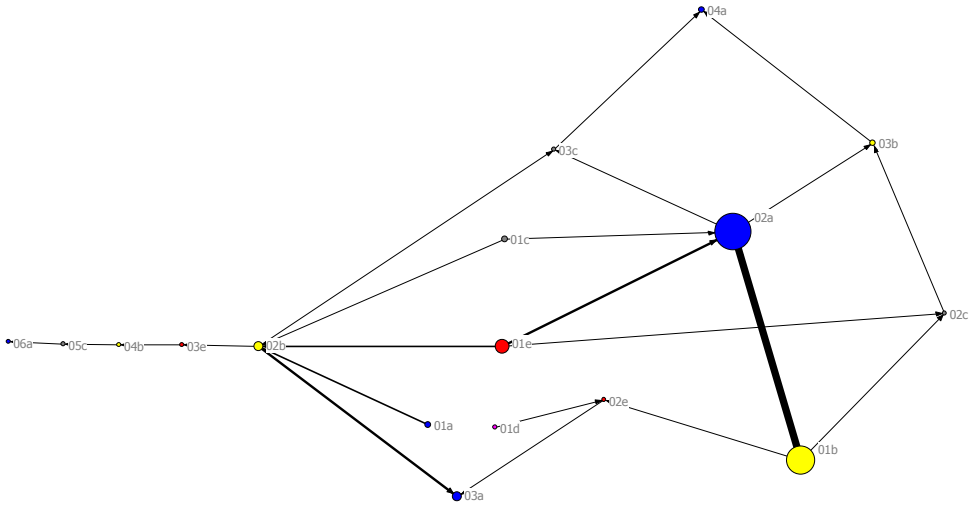
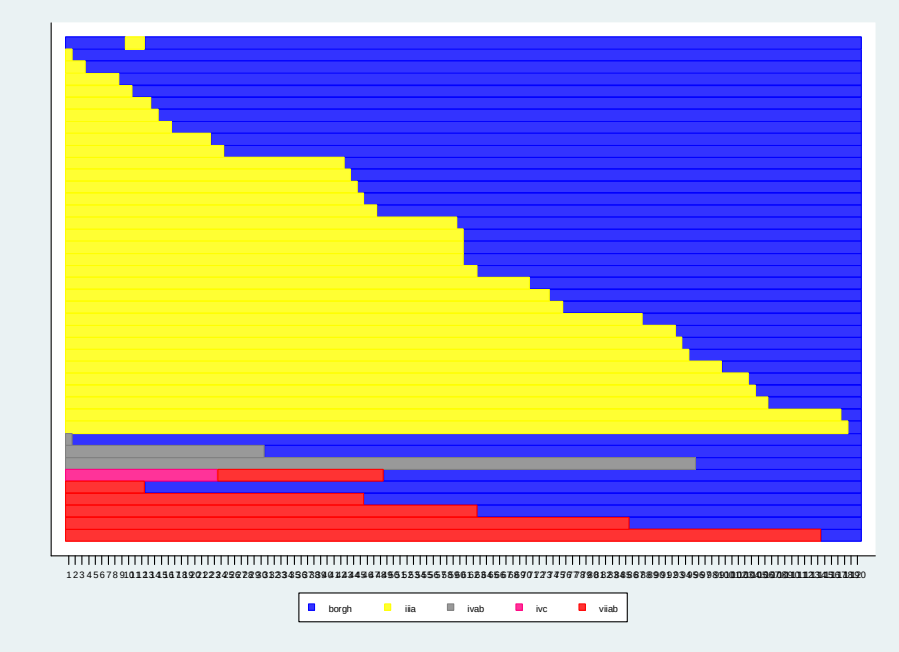
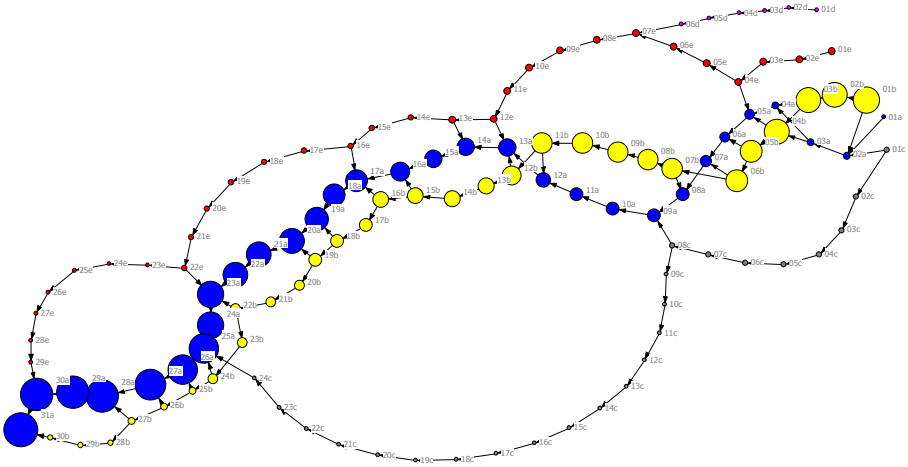


Figure A.10.0. Class careers that end in class I & II. Women. (N=42)

Sequence Index plot



Time Sequence Network



Event Sequence Network

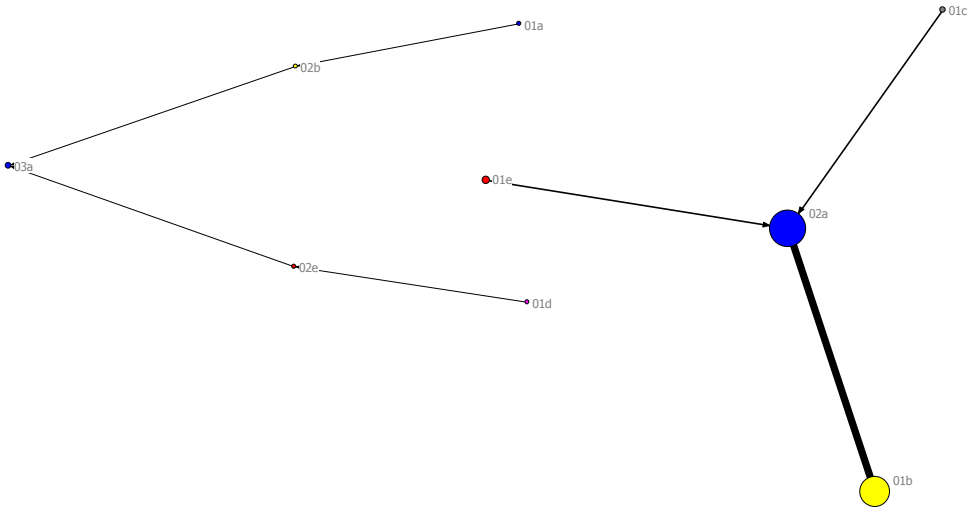
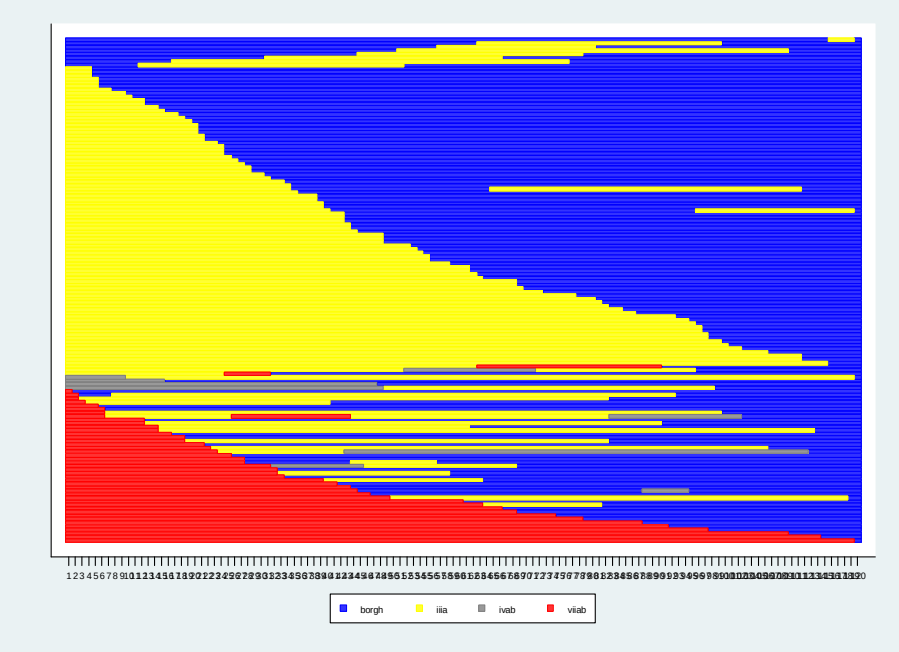
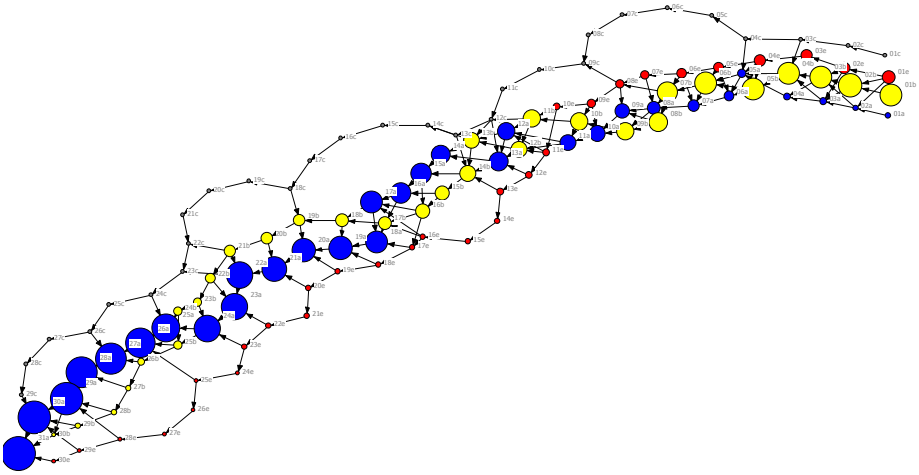


Figure A.11.0. Class careers that end in class I & II. Men (N=142)

Sequence Index plot



Time Sequence Network



Event Sequence Network

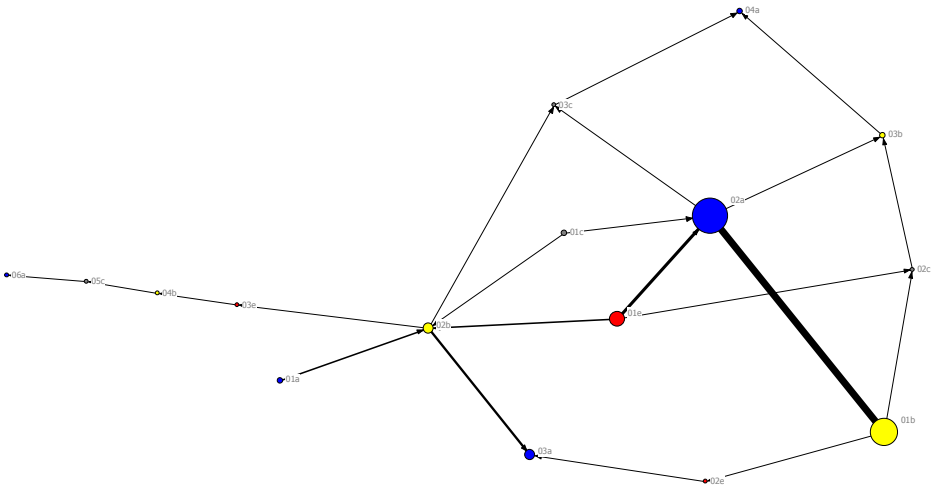


Figure A.12.0. Time sequence network of working class careers that end in class I & II.

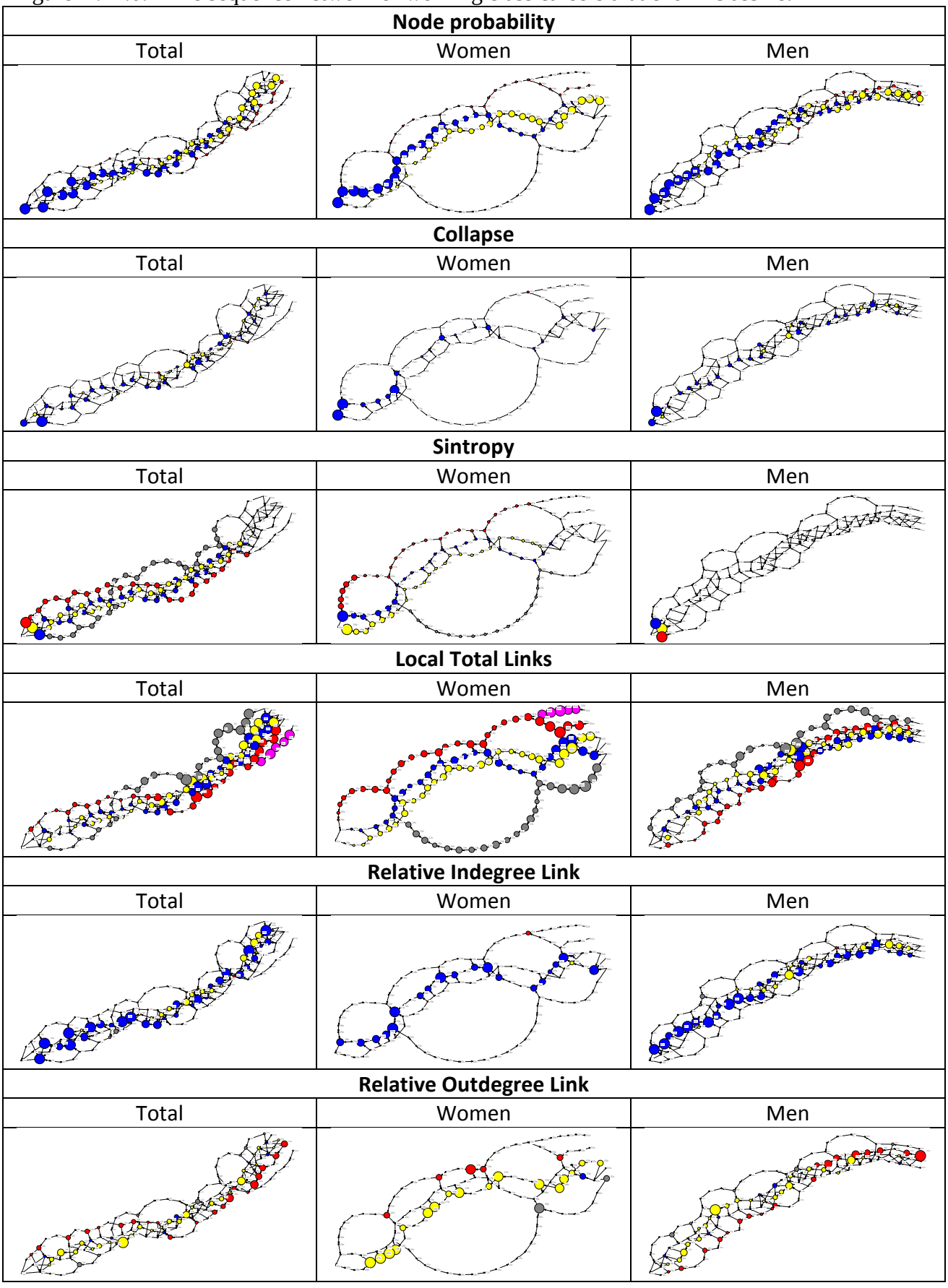
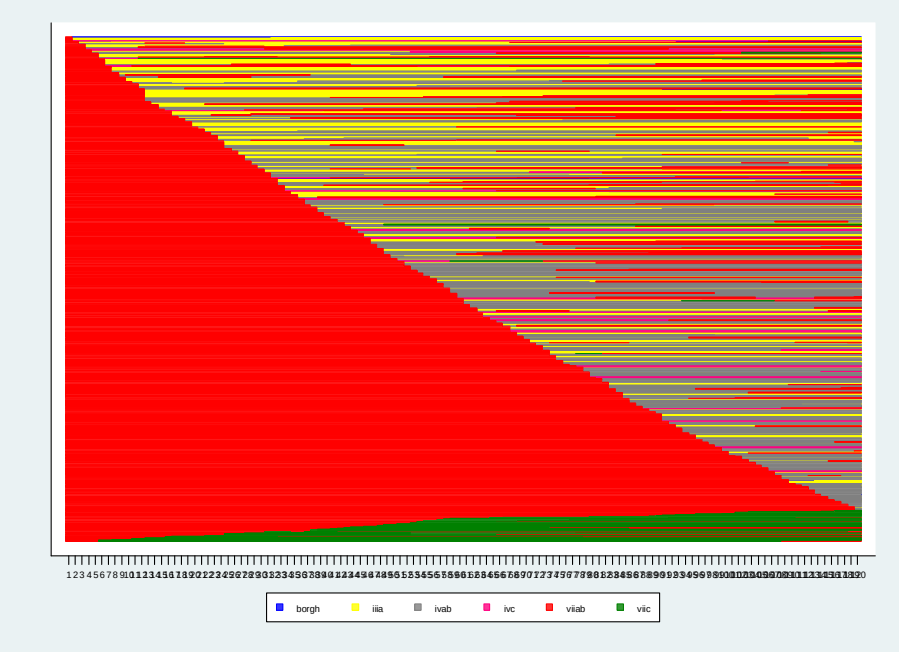
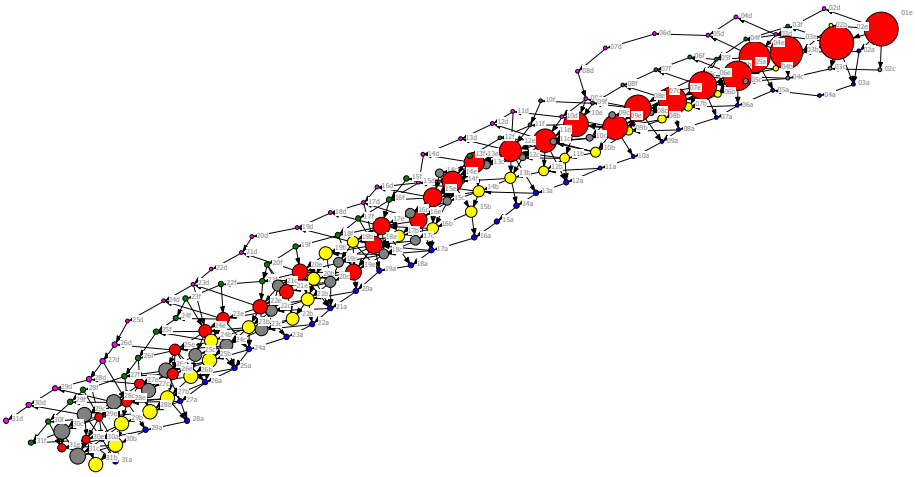


Figure A.13.0. Working class careers that begin in class IIIb+V-VI+VIIa. Total (N=880)

Sequence Index plot



Time Sequence Network



Event Sequence Network

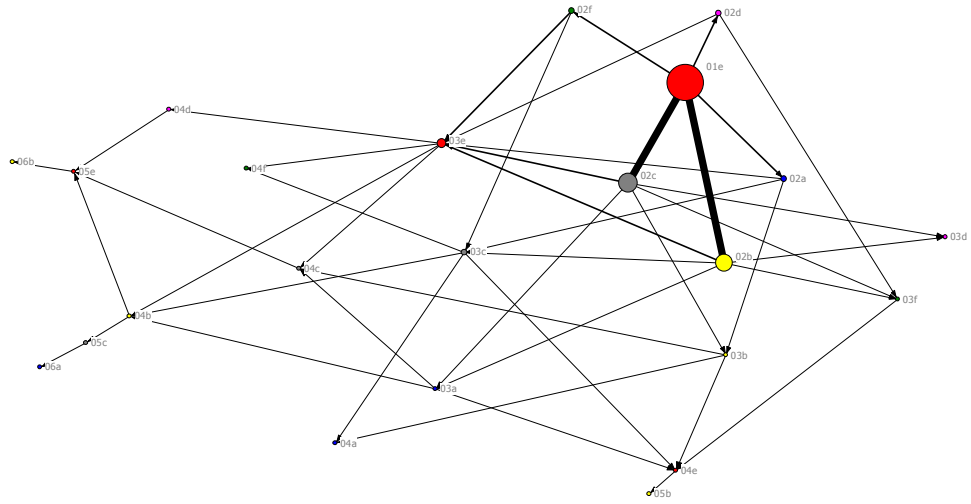
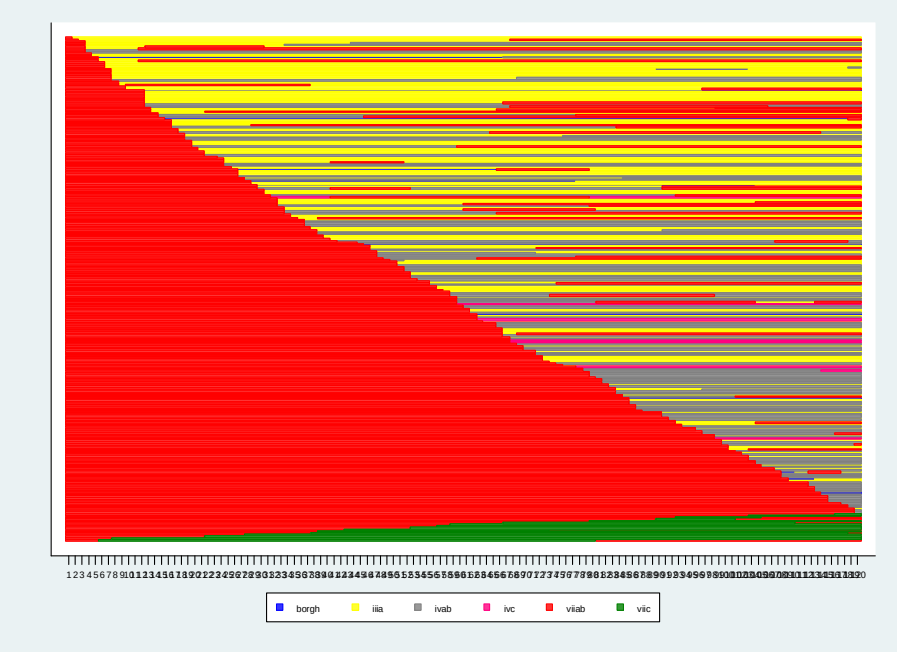
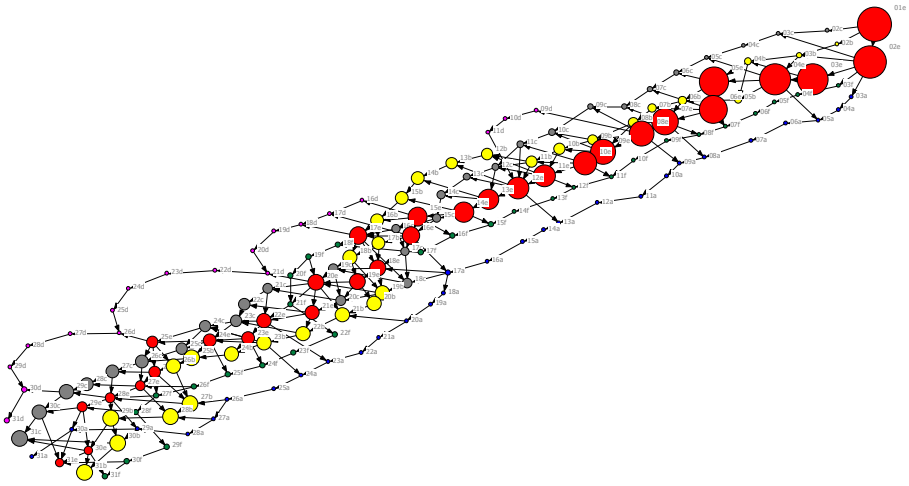


Figure A.14.0. Working class careers that begin in class IIIb+V-VI+VIIa. Women. (N=288)

Sequence Index plot



Time Sequence Network



Event Sequence Network

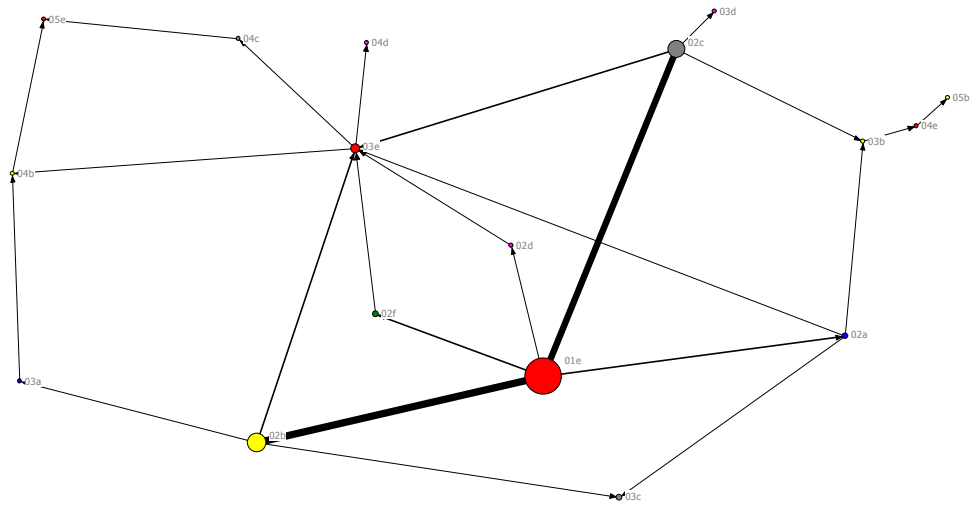
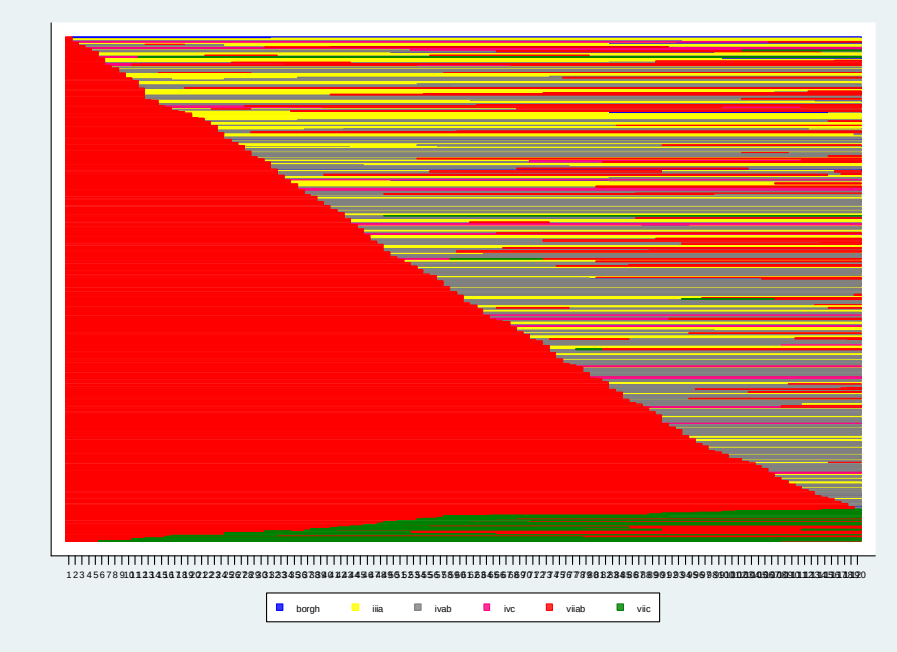
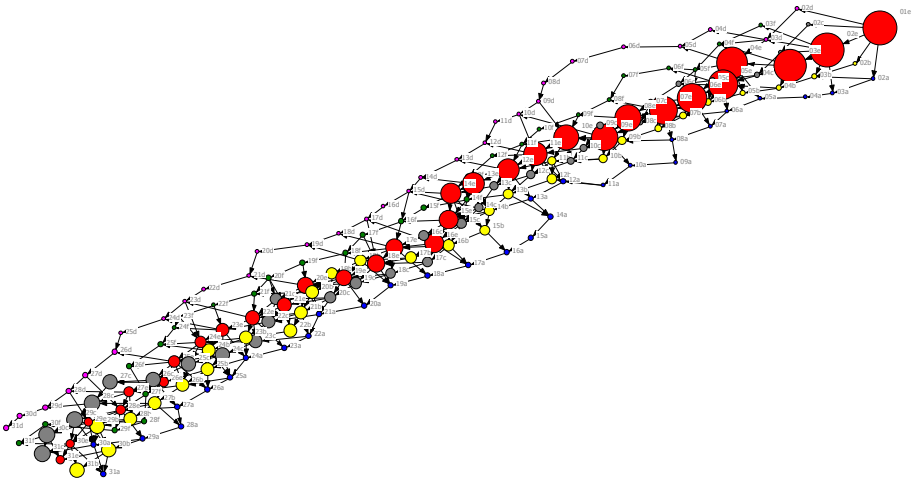


Figure A.15.0. Working class careers that begin in class IIIb+V-VI+VIIa. Men(N=592)

Sequence Index plot



Time Sequence Network



Event Sequence Network

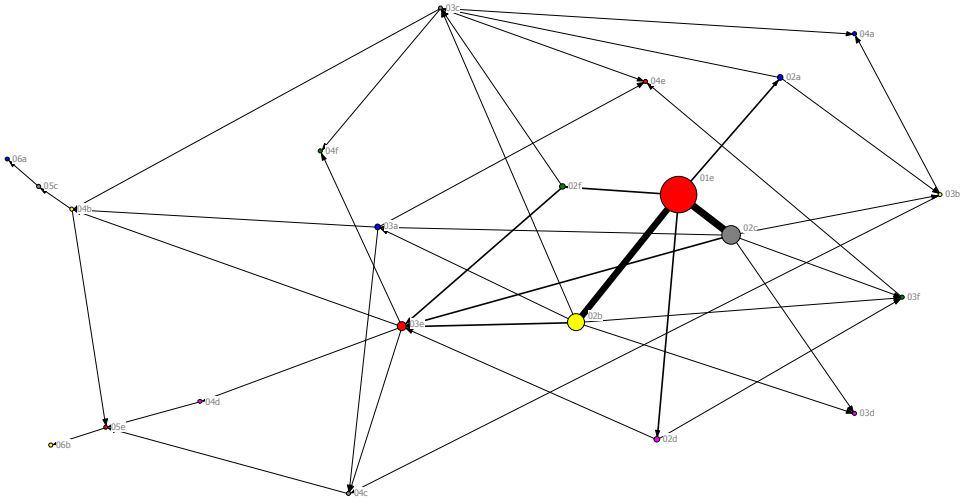
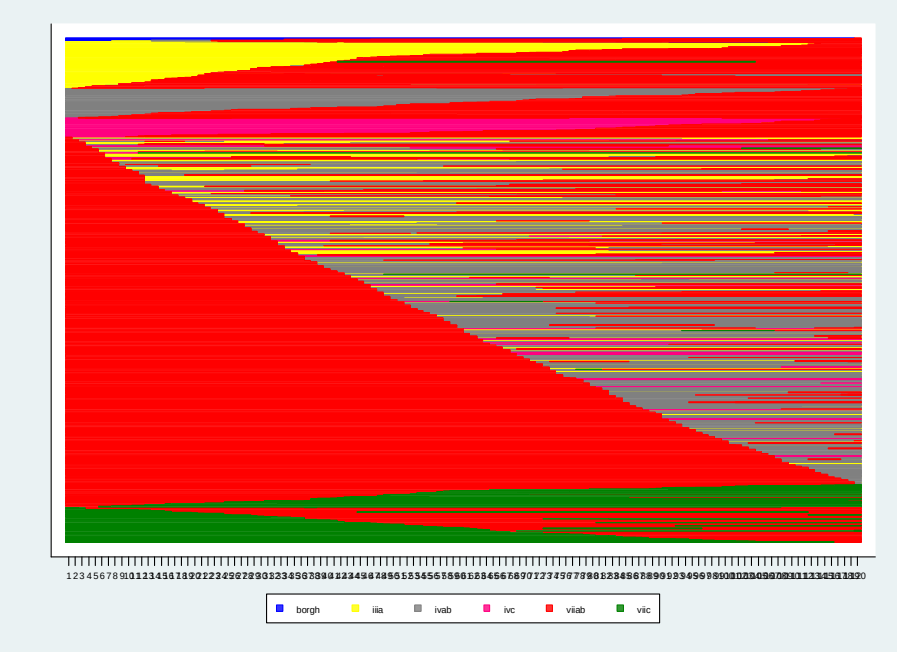
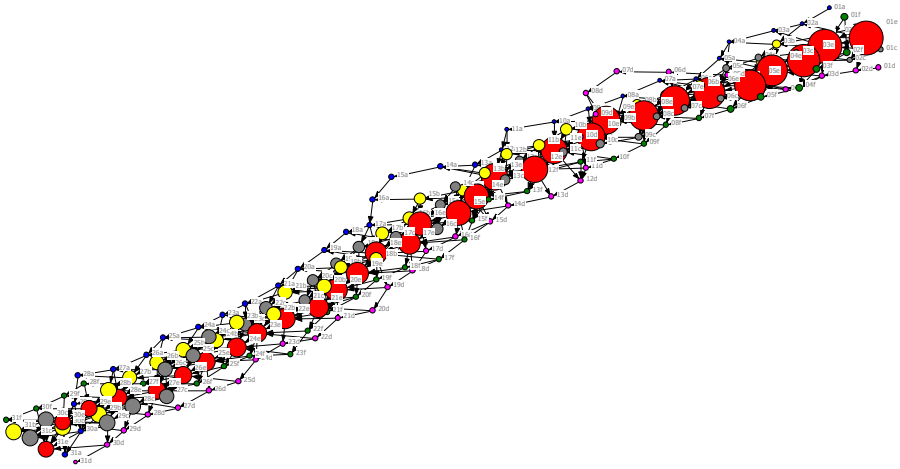


Figure A.16.0. Working class careers where at least one month has been spent in class IIIb+V-VI+VIIa. Total (N=1198)

Sequence Index plot



Time Sequence Network



Event Sequence Network

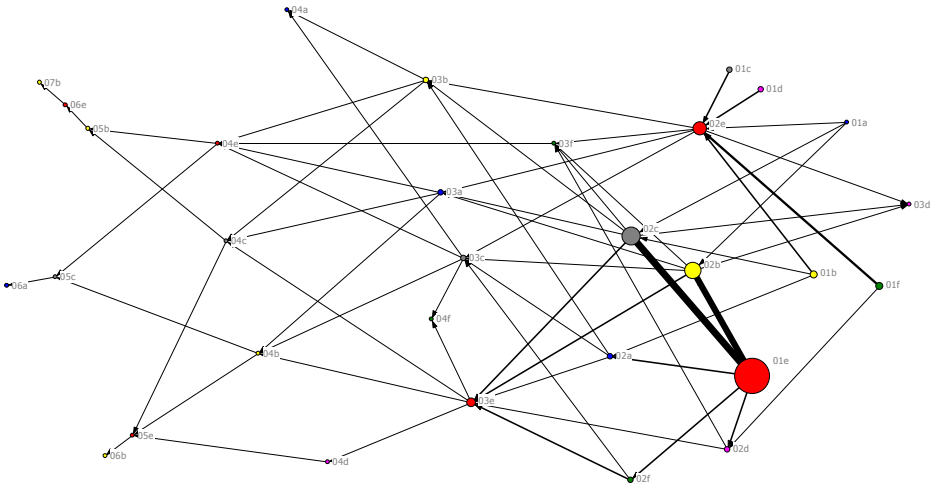
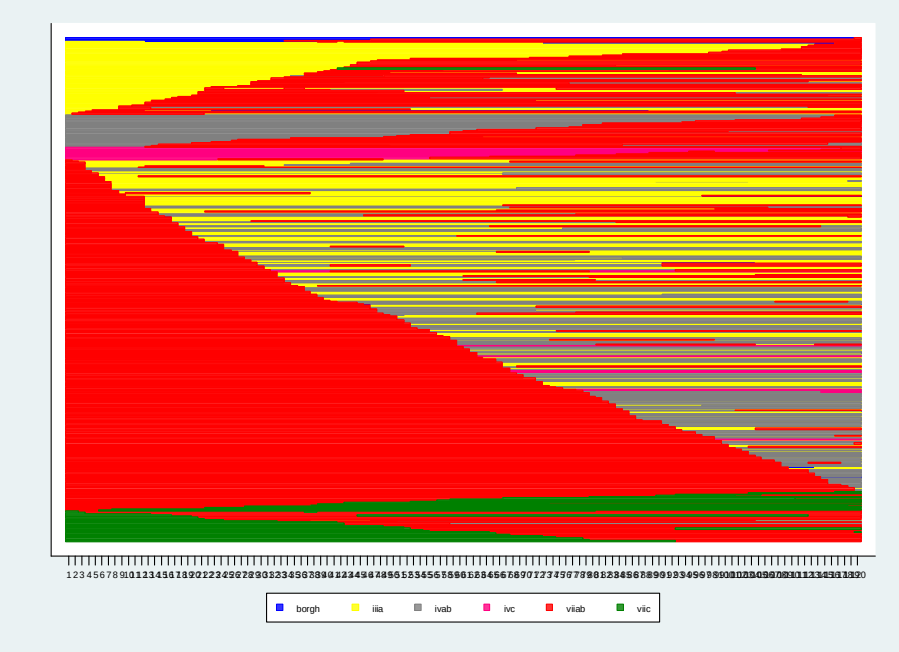
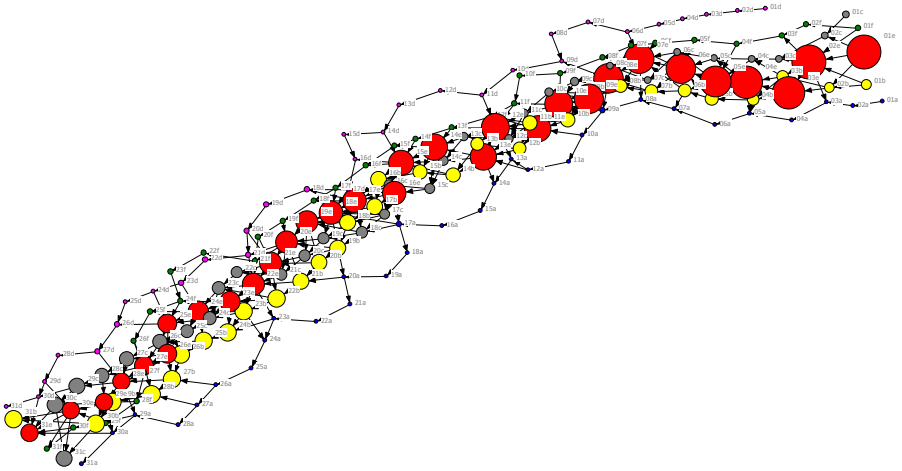


Figure A.17.0. Working class careers where at least one month has been spent in class IIIb+V-VI+VIIa. Women(N=413)

Sequence Index plot



Time Sequence Network



Event Sequence Network

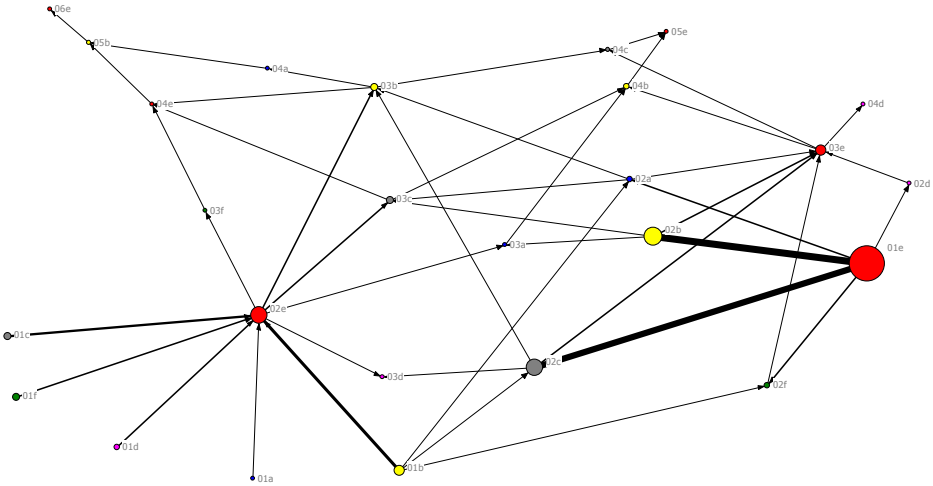
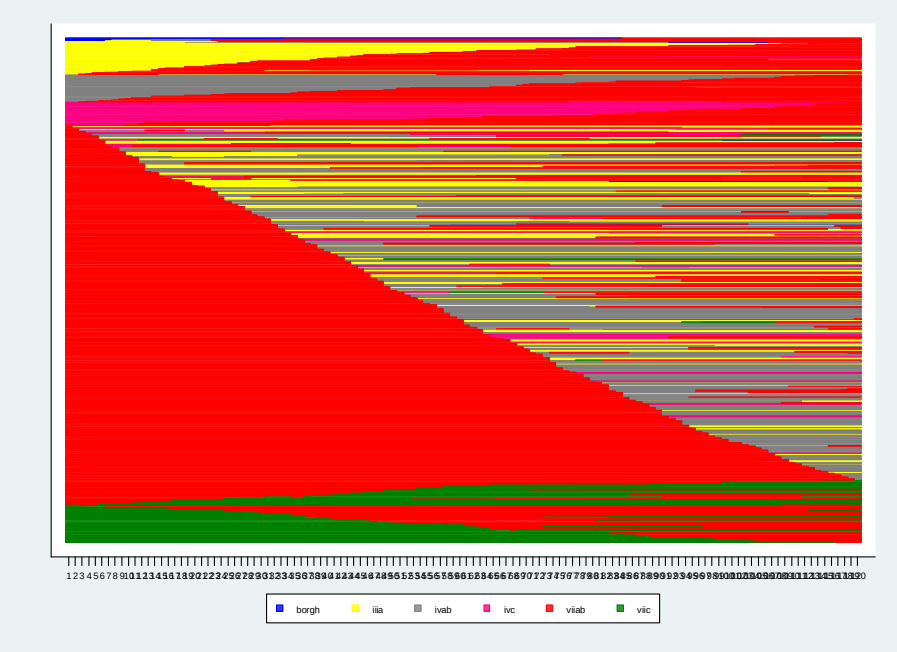
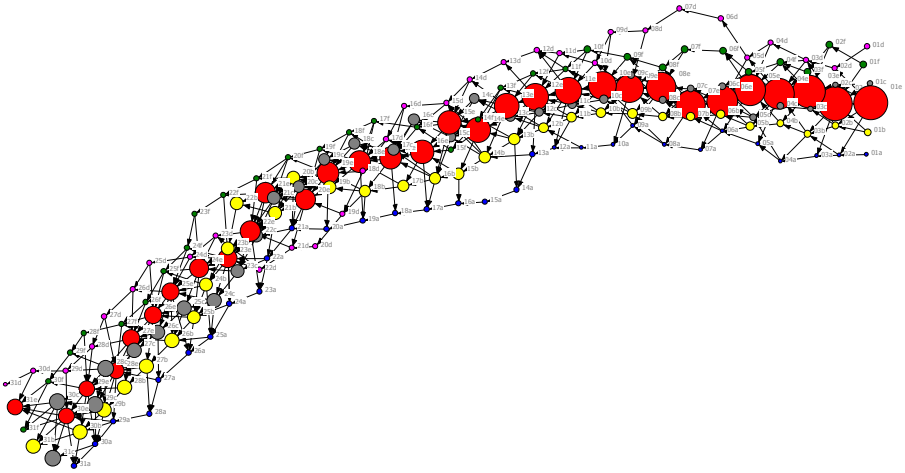


Figure A.18.0. Working class careers where at least one month has been spent in class IIIb+V-VI+VIIa. Men (N=785)

Sequence Index plot



Time Sequence Network



Event Sequence Network

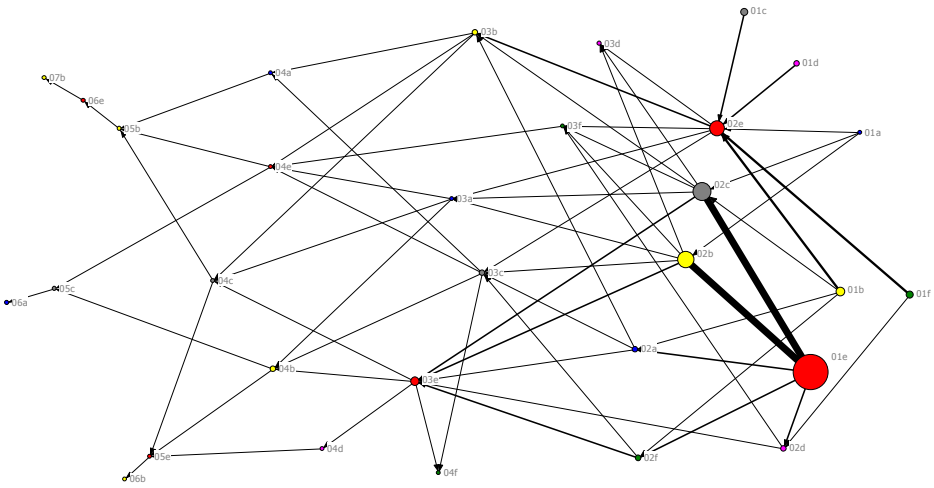
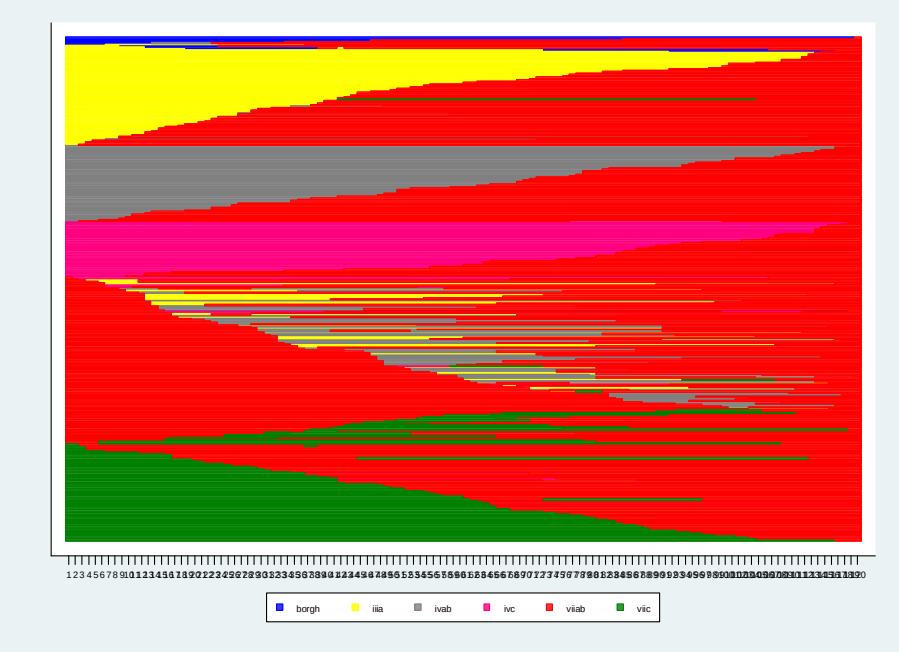
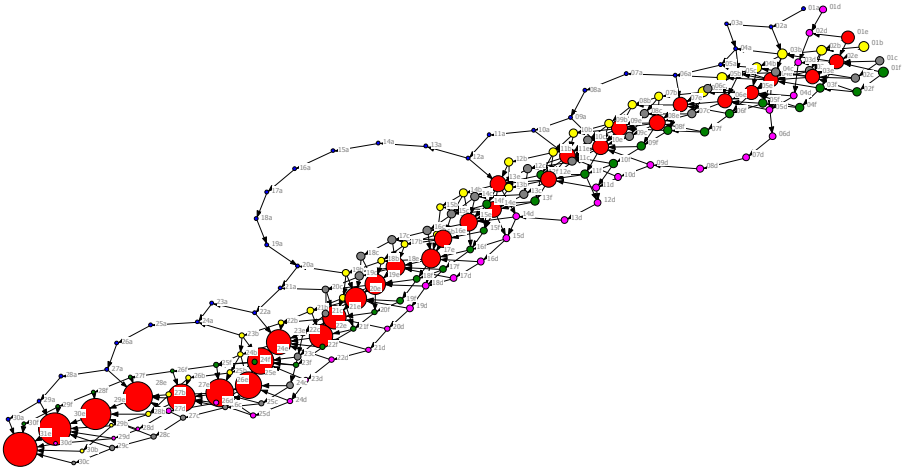


Figure A.19.0. Class careers that end in class IIIb+V-VI+VIIa. Total (N=368)

Sequence Index plot



Time Sequence Network



Event Sequence Network

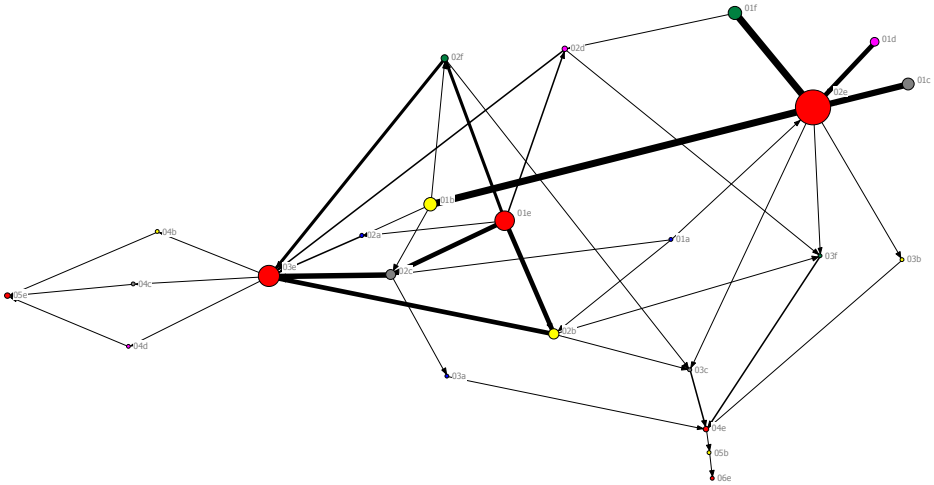
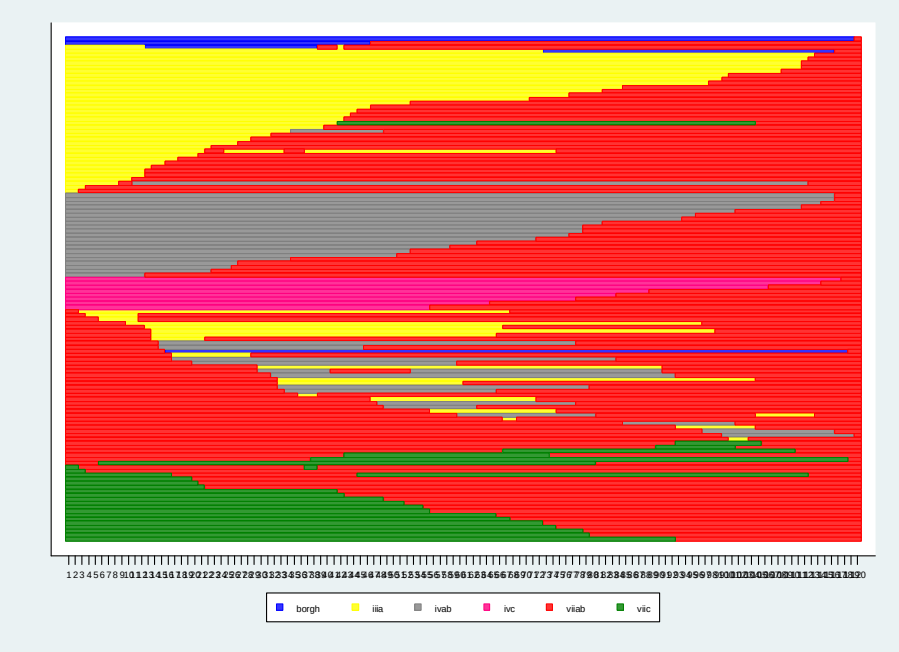
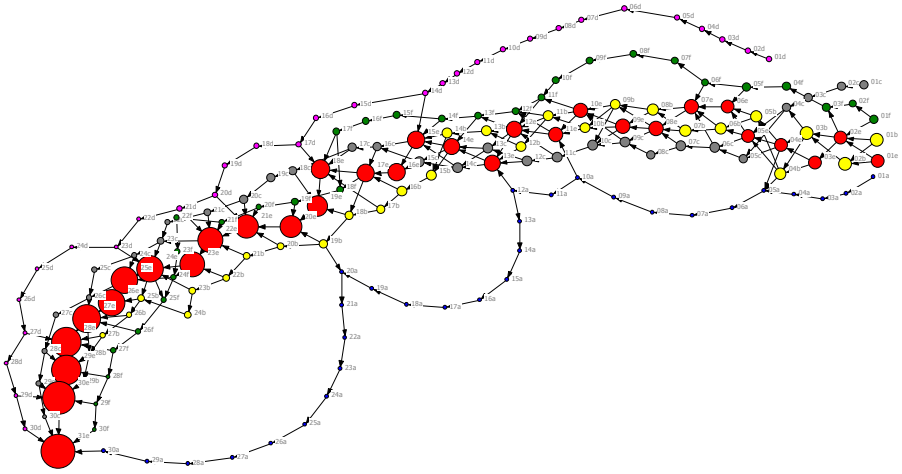


Figure A.20.0. Class careers that end in class IIIb+V-VI+VIIa. Women(N=126)

Sequence Index plot



Time Sequence Network



Event Sequence Network

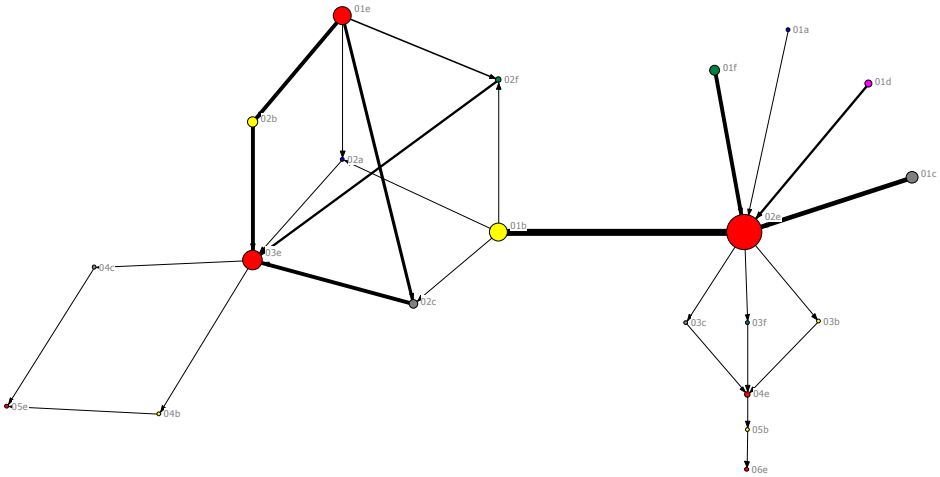
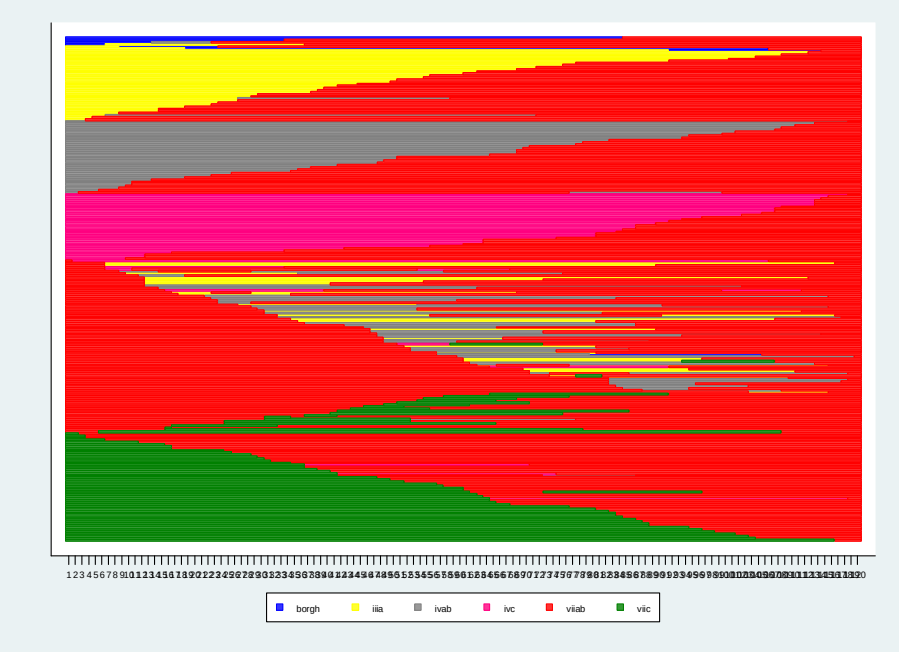
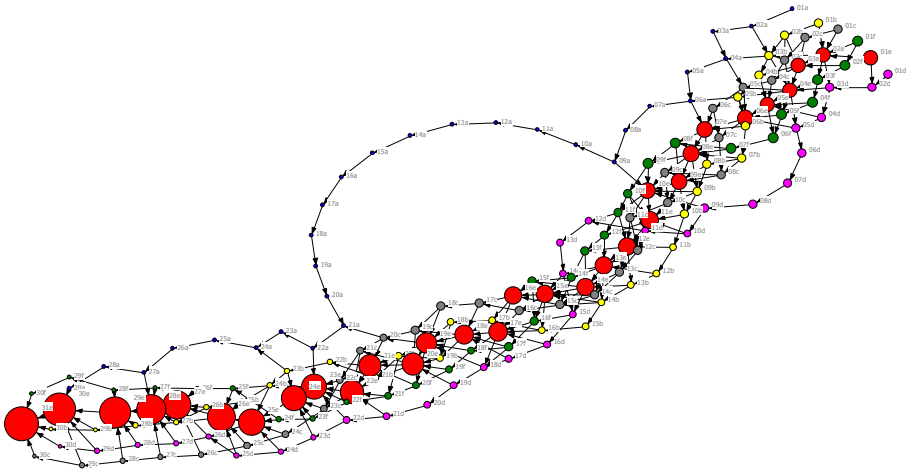


Figure A.21.0. Class careers that end in class IIIb+V-VI+VIIa. Men (N=242)

Sequence Index plot



Time Sequence Network



Event Sequence Network

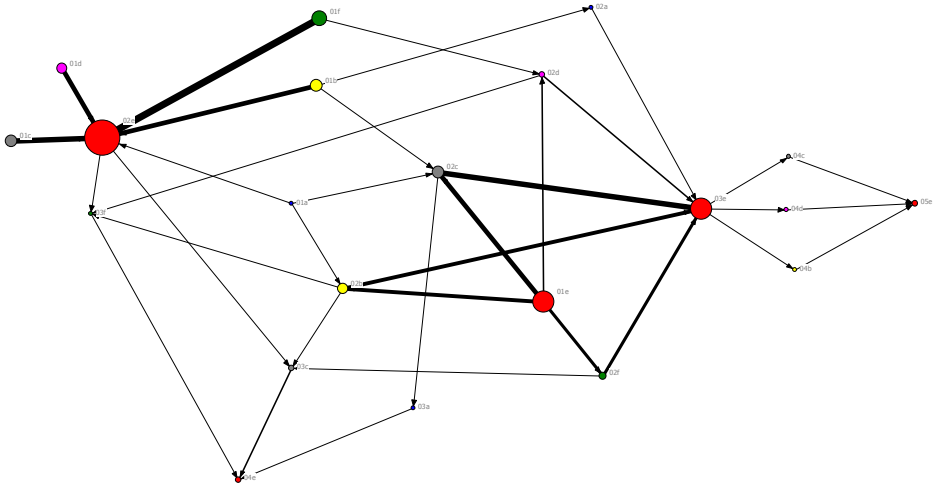
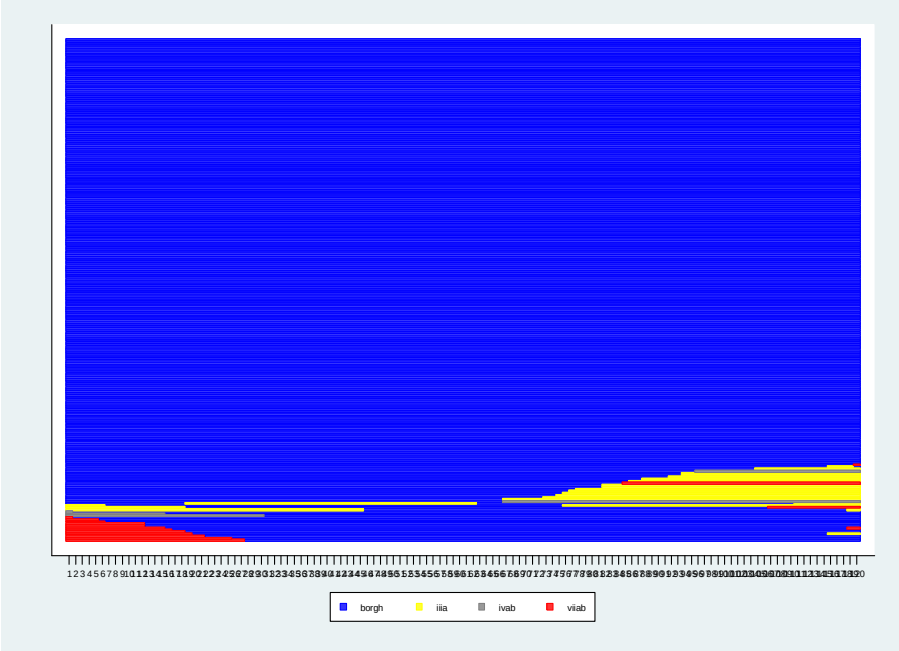
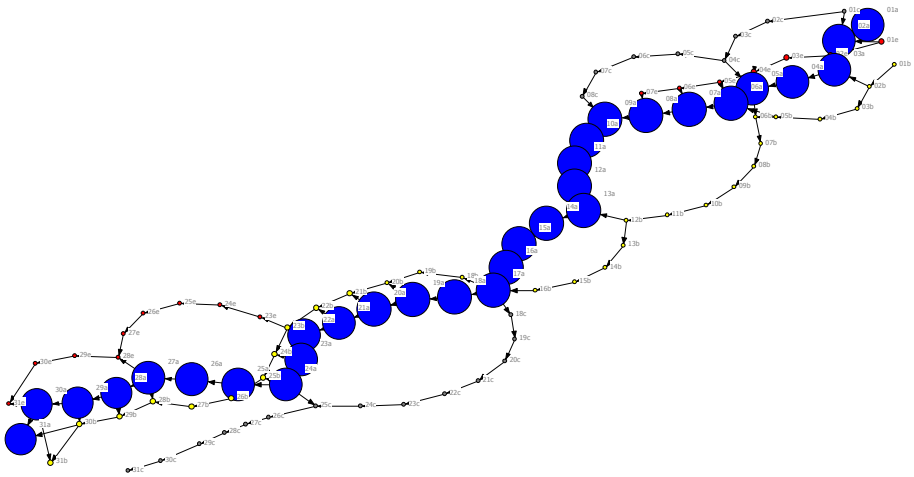


Figure A.22.0. Immobility pattern I+II: Cluster a (N=247)

Sequence Index plot



Time Sequence Network



Event Sequence Network

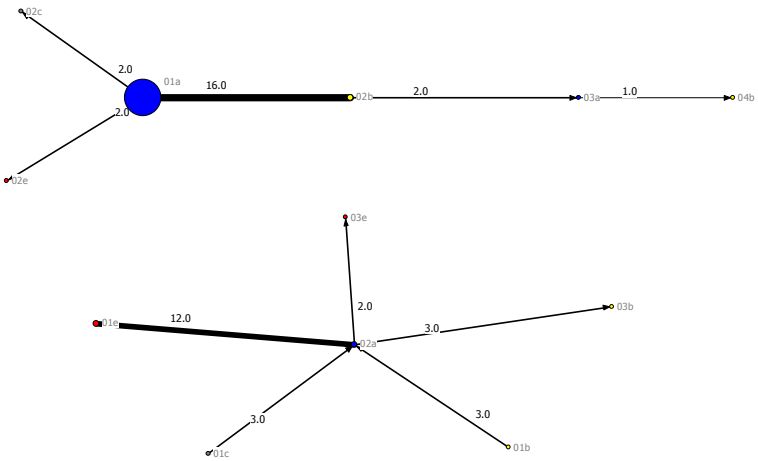
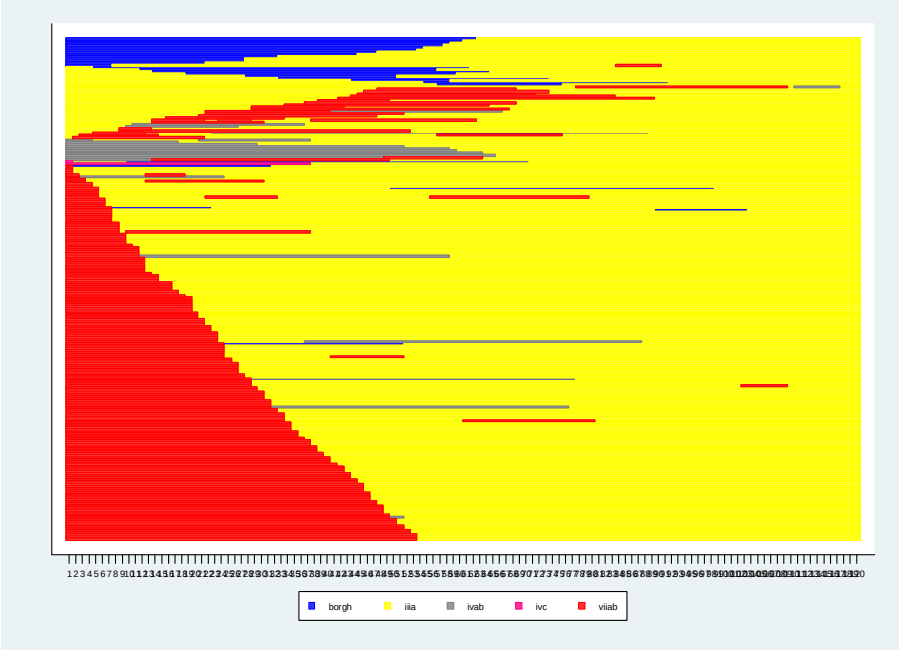
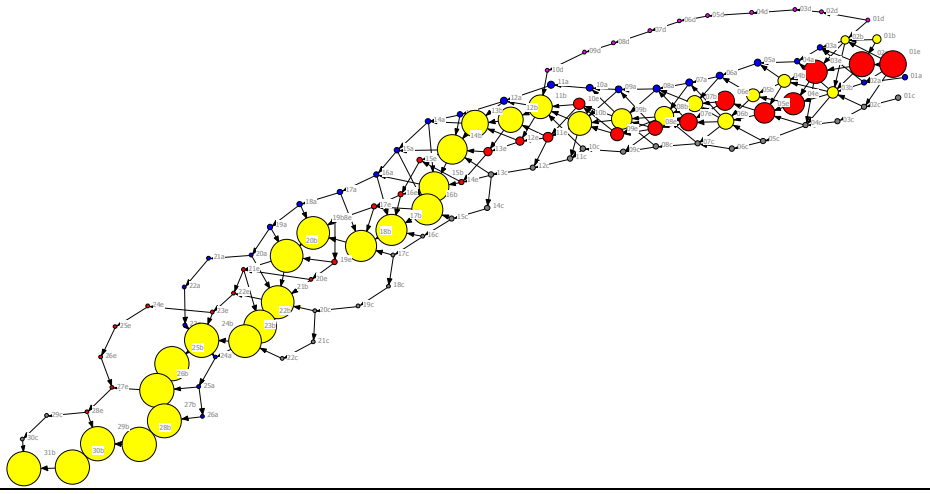


Figure A.23.0. Upwards mobility patterns: IIIb+V-VI+VIIa→IIIa & IIIa→IIIb+V-VI+VIIa→IIIa (fast); Cluster I (N=229)

Sequence Index plot



Time Sequence Network



Event Sequence Network

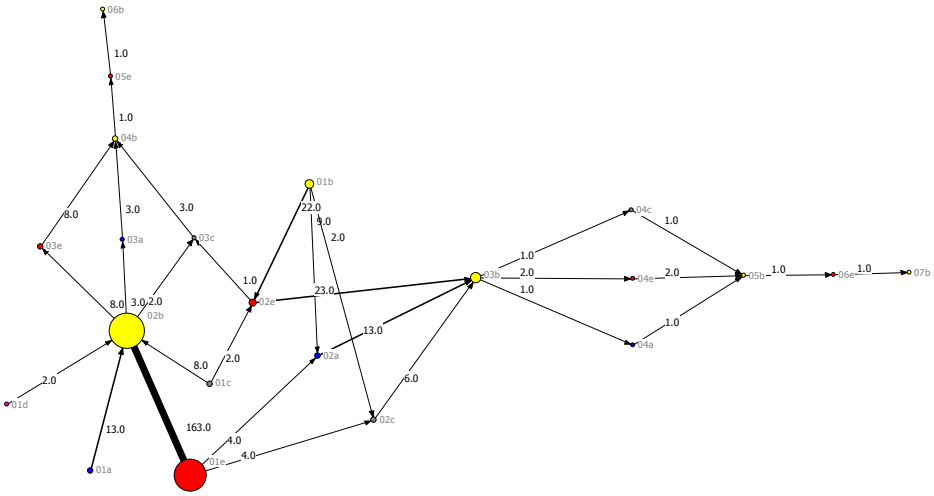
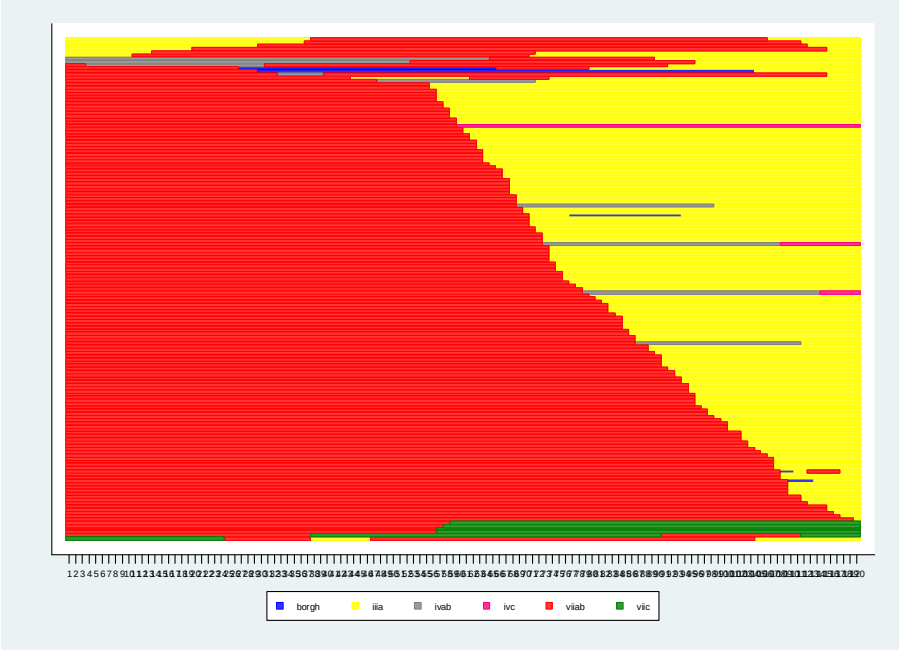
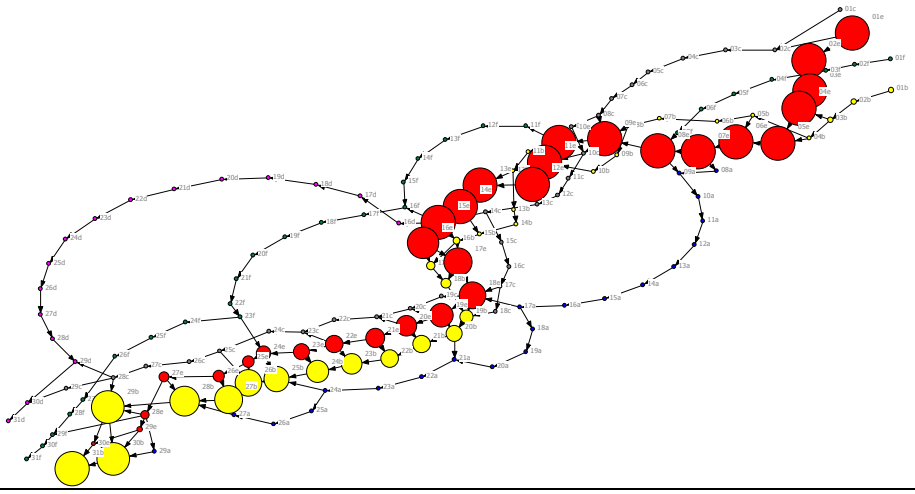


Figure A.24.0. Upwards mobility patterns: IIIb+V-VI+VIIa→IIIa & IIIa→IIIb+V-VI+VIIa→IIIa (slow); Cluster m (157)

Sequence Index plot



Time Sequence Network



Event Sequence Network

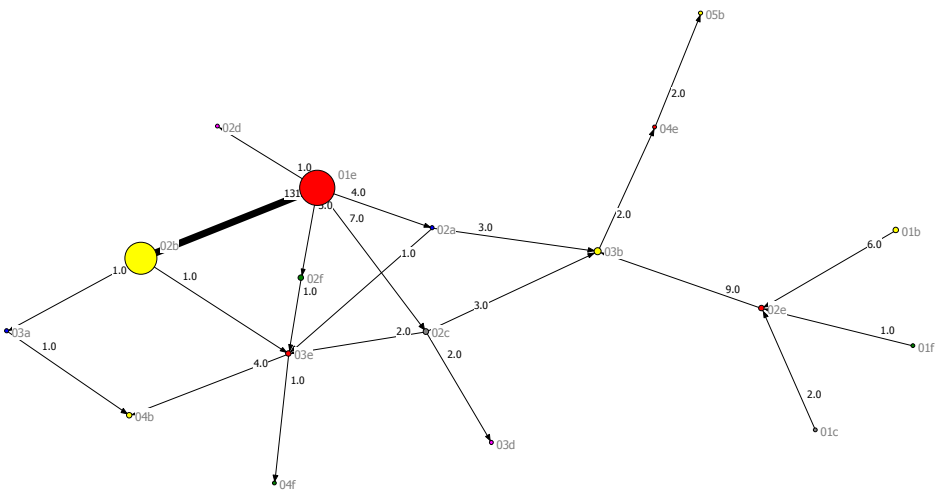
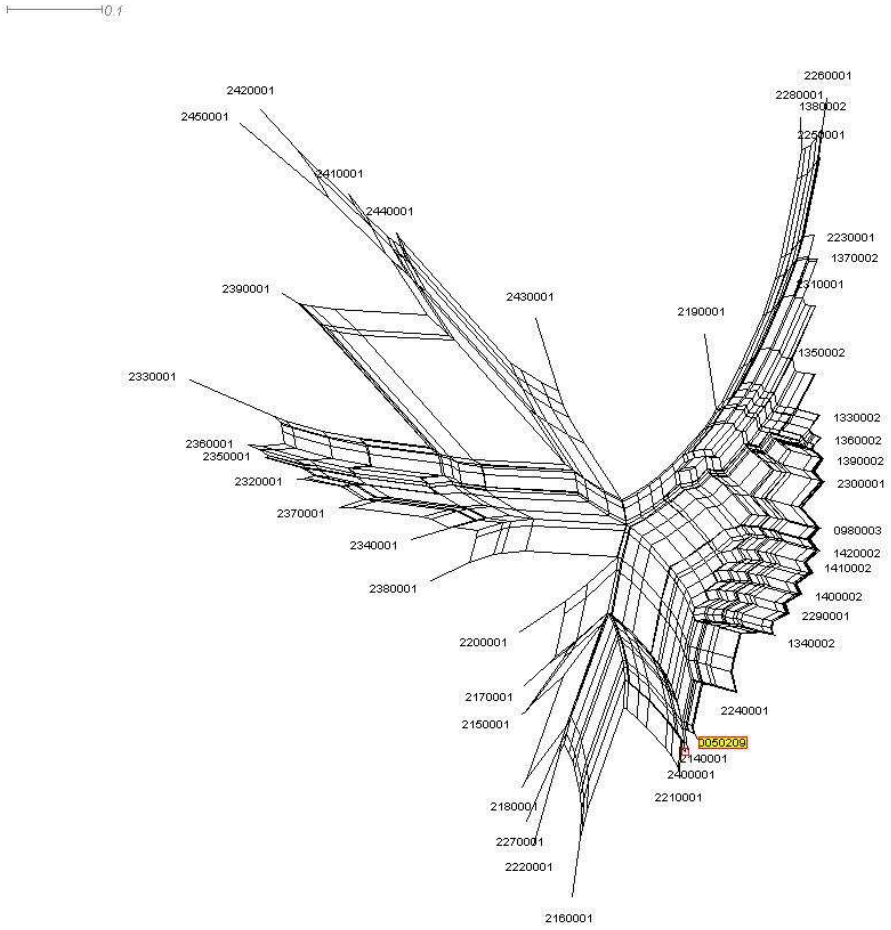
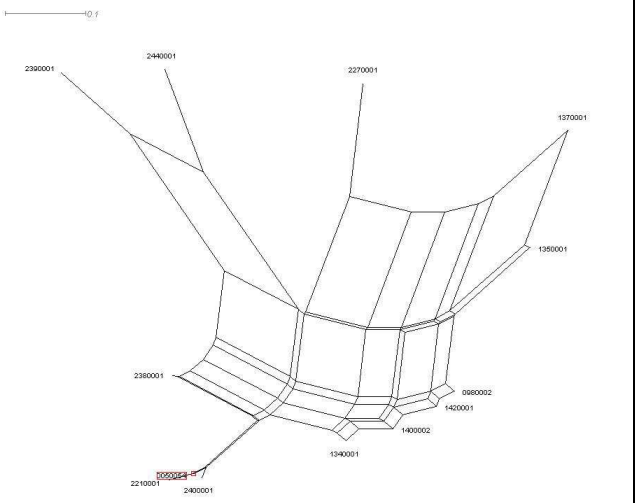


Figure A.25.0. Phylogenetic network of class careers that begin in class I & II. On total and by gender

Class I & II Total



Class I & II Women



Class I & II Men

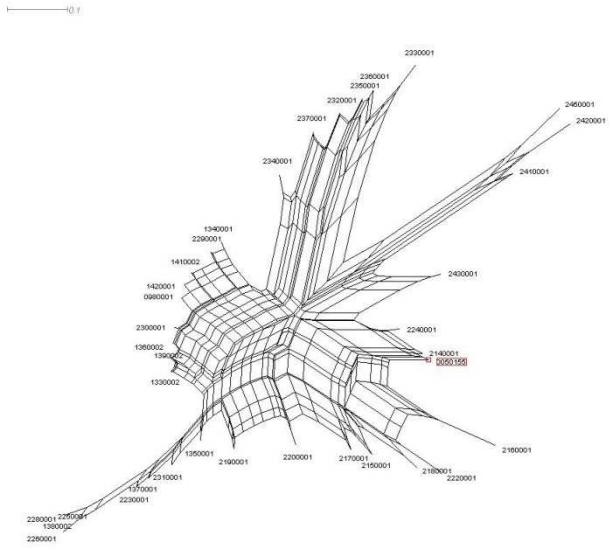
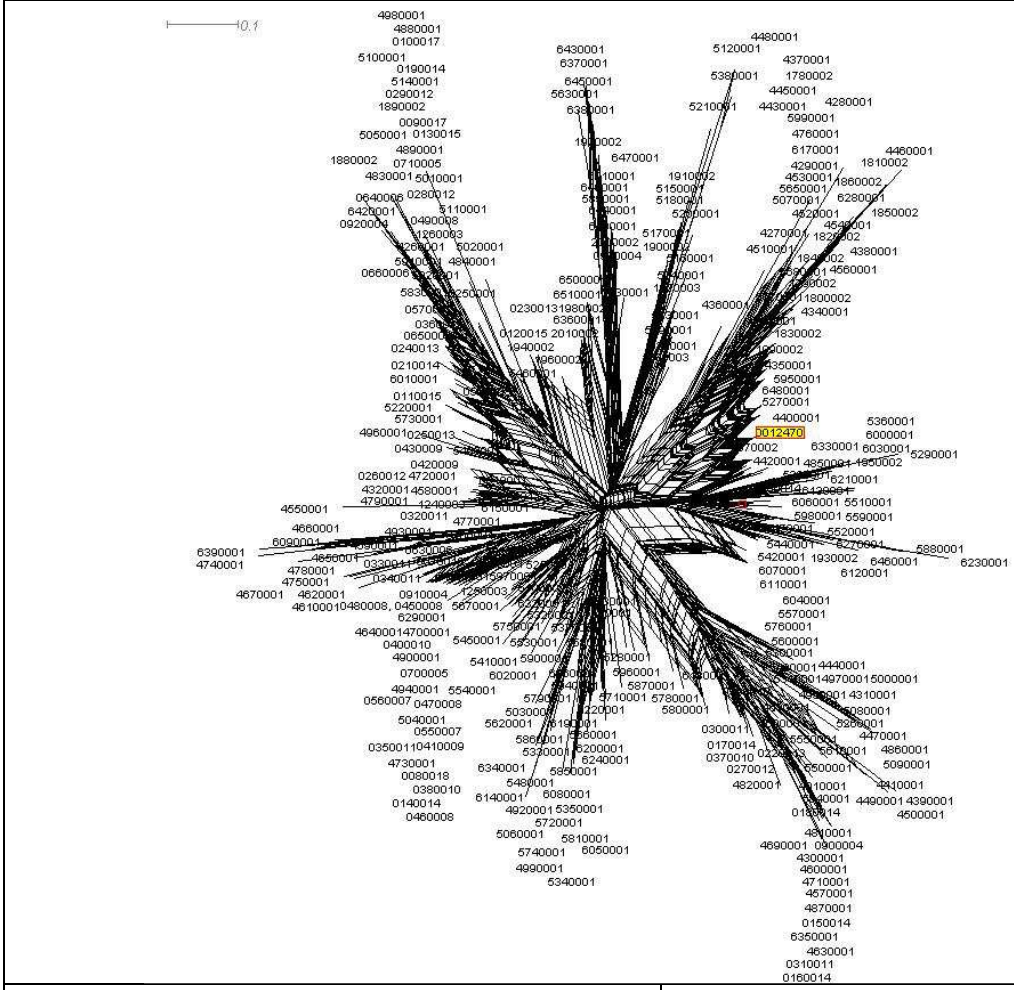
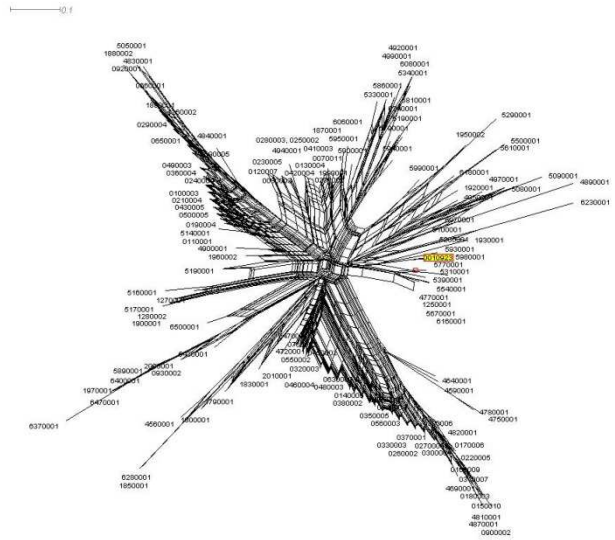


Figure A.26.0. Phylogenetic network of class careers that begin in class IIIb+V-VI+VIIa. On Total and by gender

Class IIIb+V-VI+VIIa Total



Class IIIb+V-VI+VIIa Women



Class IIIb+V-VI+VIIa Men

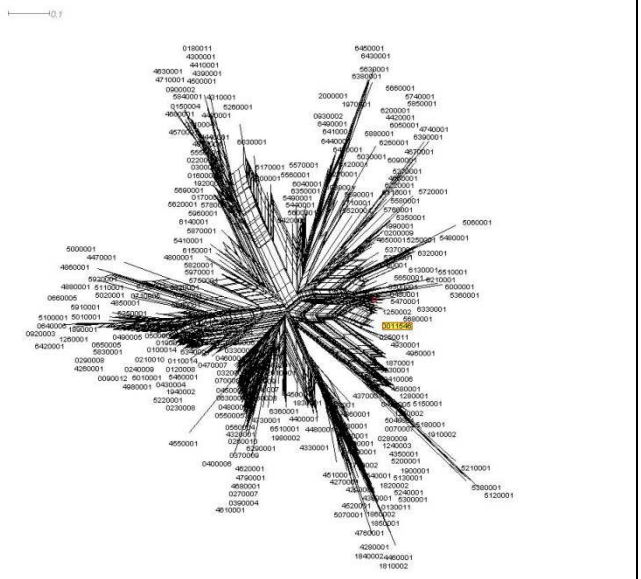
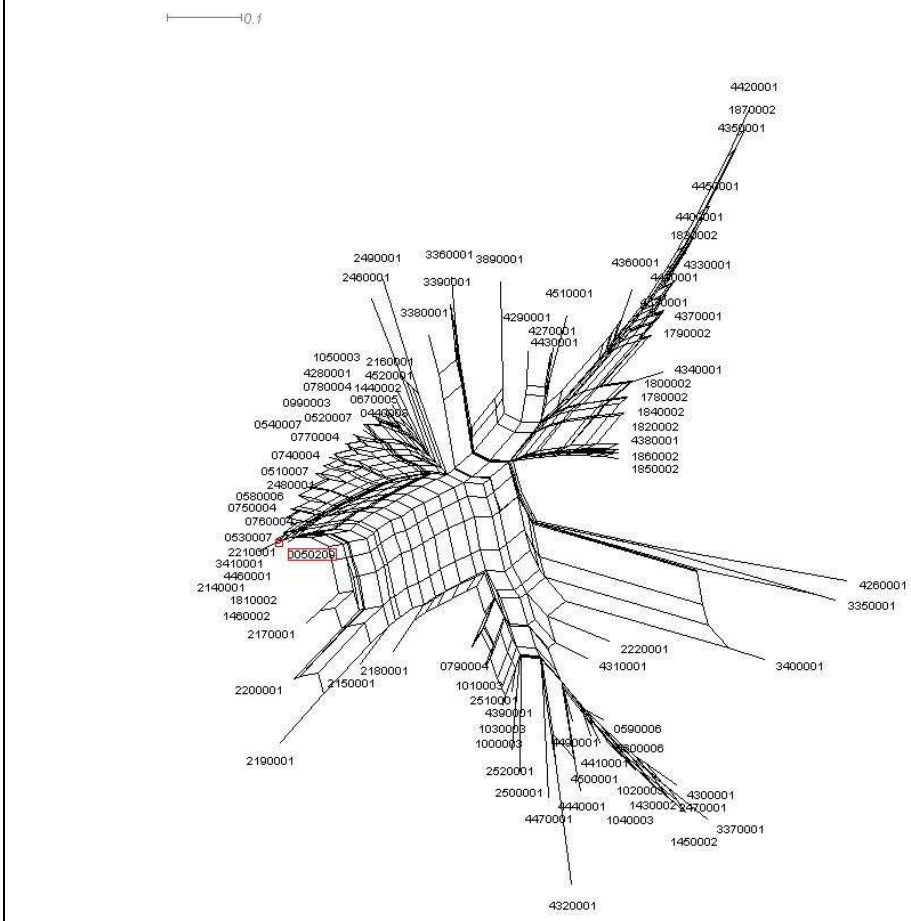


Figure A.27.0. Phylogenetic network of class careers that end in class I & II. On Total and by gender

Class I & II Tot



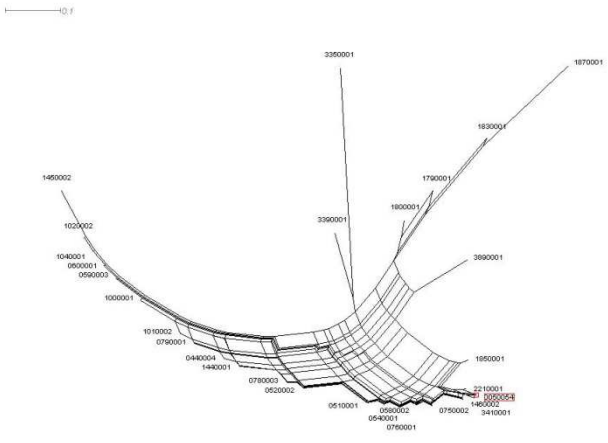
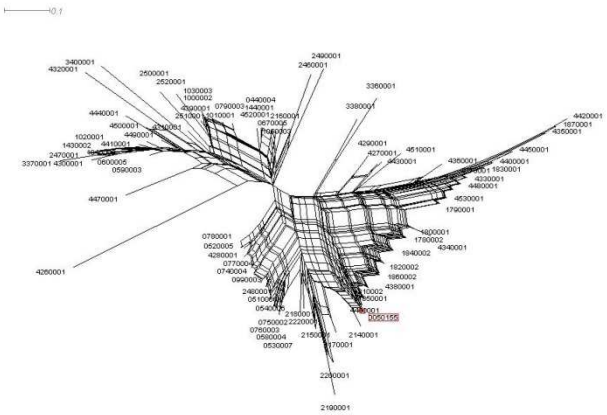
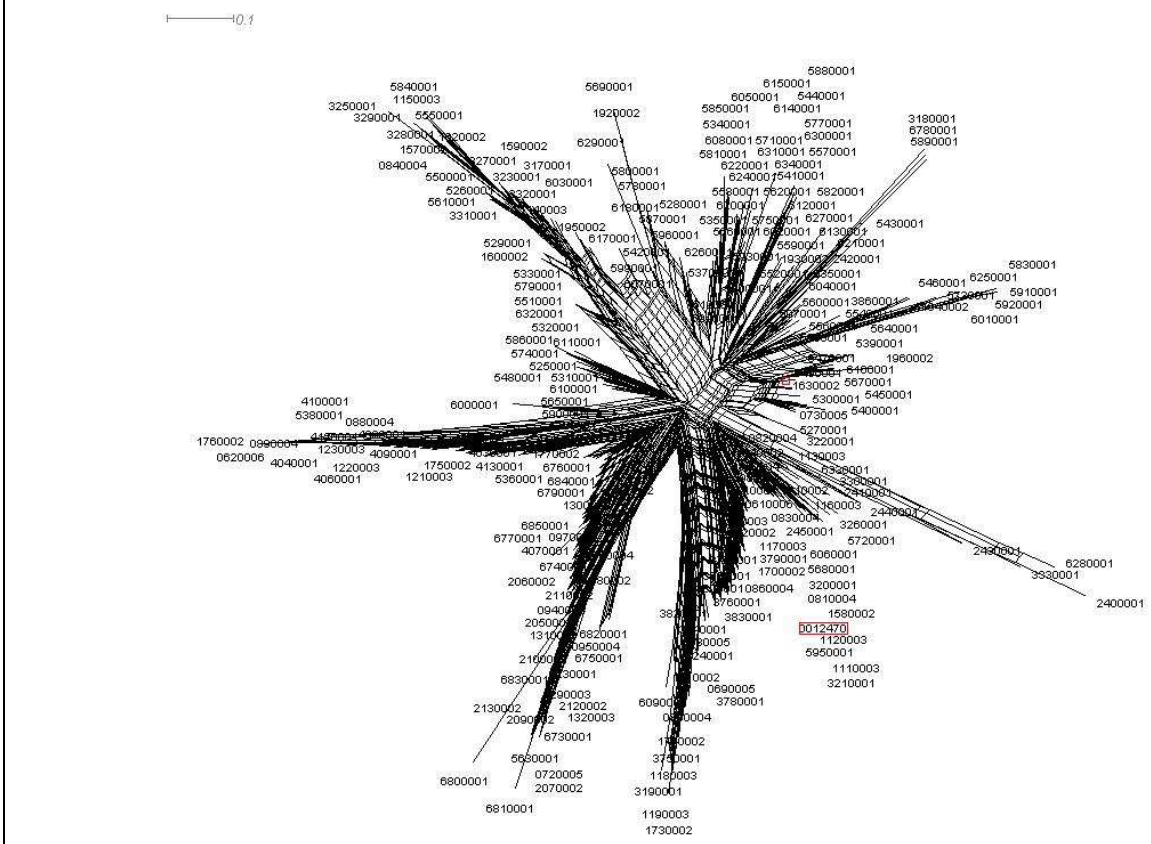
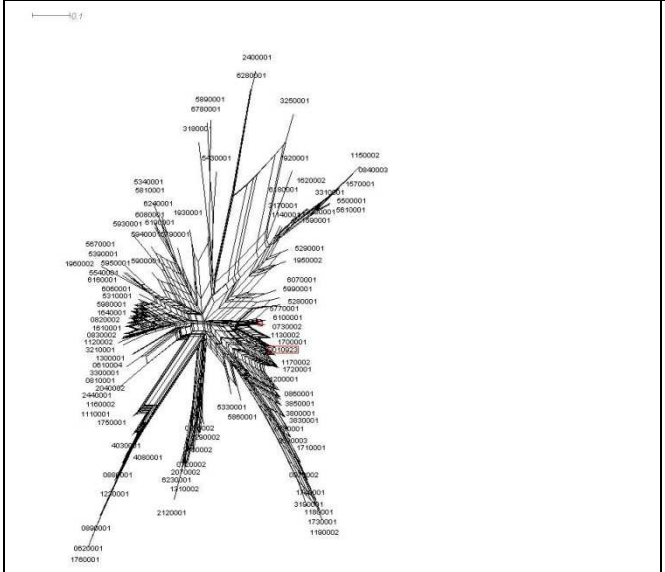
Class I & II Women	Class I & II Men
	

Figure A.28.0. Phylogenetic network of class careers that end in class IIIb+V-VI+VIIa. On Total and by gender

Class IIIb+V-VI+VIIa Total



Class IIIb+V-VI+VIIa Women



Class IIIb+V-VI+VIIa Men