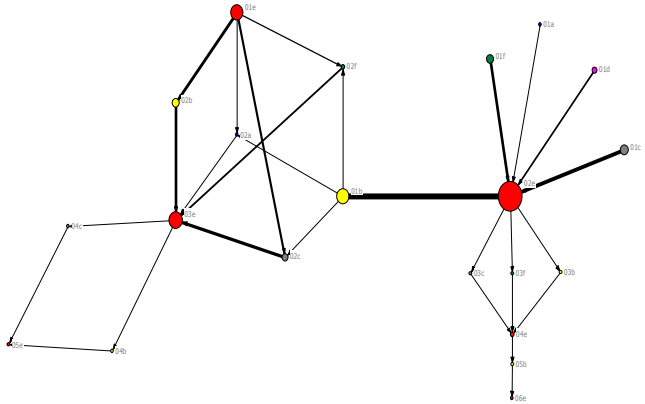
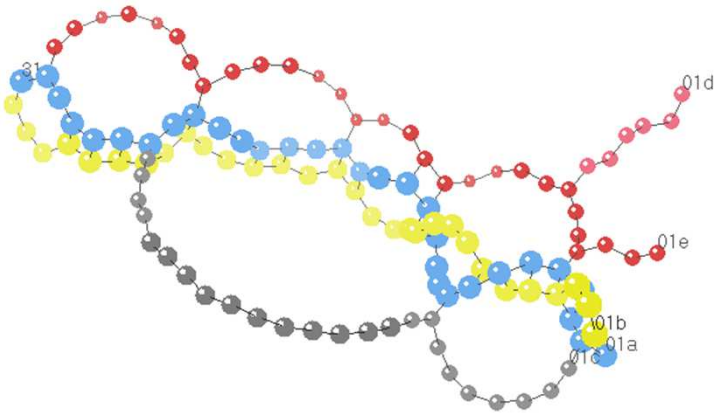
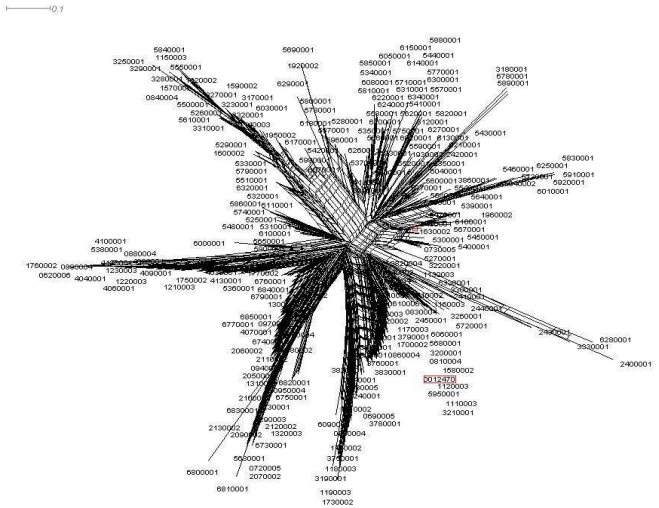
Sequence analysis and Network analysis:

An attempt to represent and study sequences by using NetDraw

By
Ivano Bison







LaCOSA

Lausanne Conference On Sequence Analysis

University of Lausanne,

June 6th-8th 2012

The purpose of the paper is to explores alternative ways:

1) to visualize sequences.

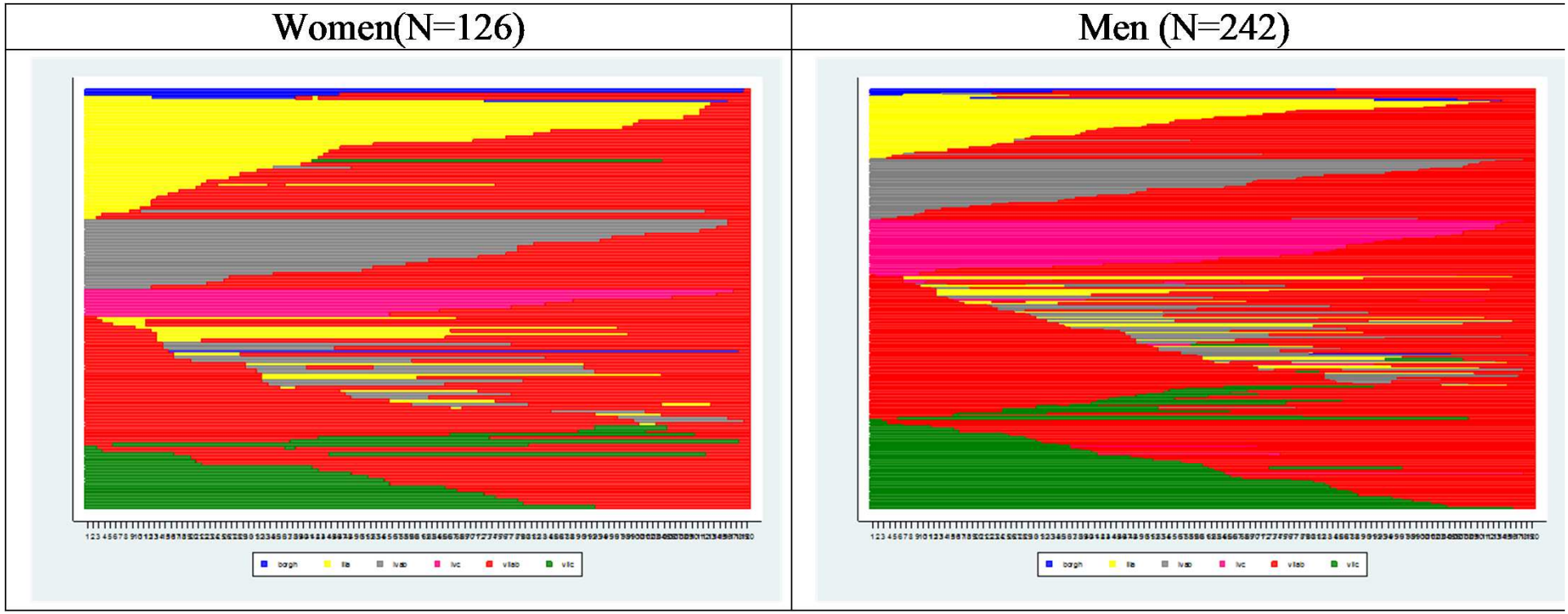
2) to extract information contained in the sequences.

The aim is to find new perspectives from which to study sequences.

The overall goal is tries to open the doors to new ways of identifying potential patterns or underlying structures.

For instance, suppose we want to identify graphically whether there are substantial differences in class careers between men and women.

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



except for differences at the beginning of their careers

the career patterns of men and women who end in

the working class after ten years are similar

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?

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 acted to different mechanisms that generate patterns that only apparently are similar but, in fact, are substantially different as regards timing and shape?

How to bring out this common pattern?

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'*.

Back to origins

Back to observe how careers develop over time.
Start again to take into account their dynamic evolution.

It is to **give physical form** to sequences and their underlying generative processes.

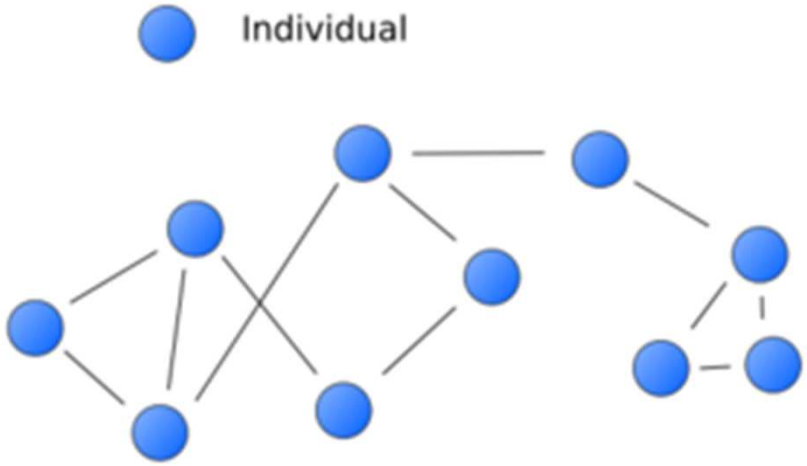
It is to **transform sequences into objects to explore** like a DNA chain or to follow like a Google map.

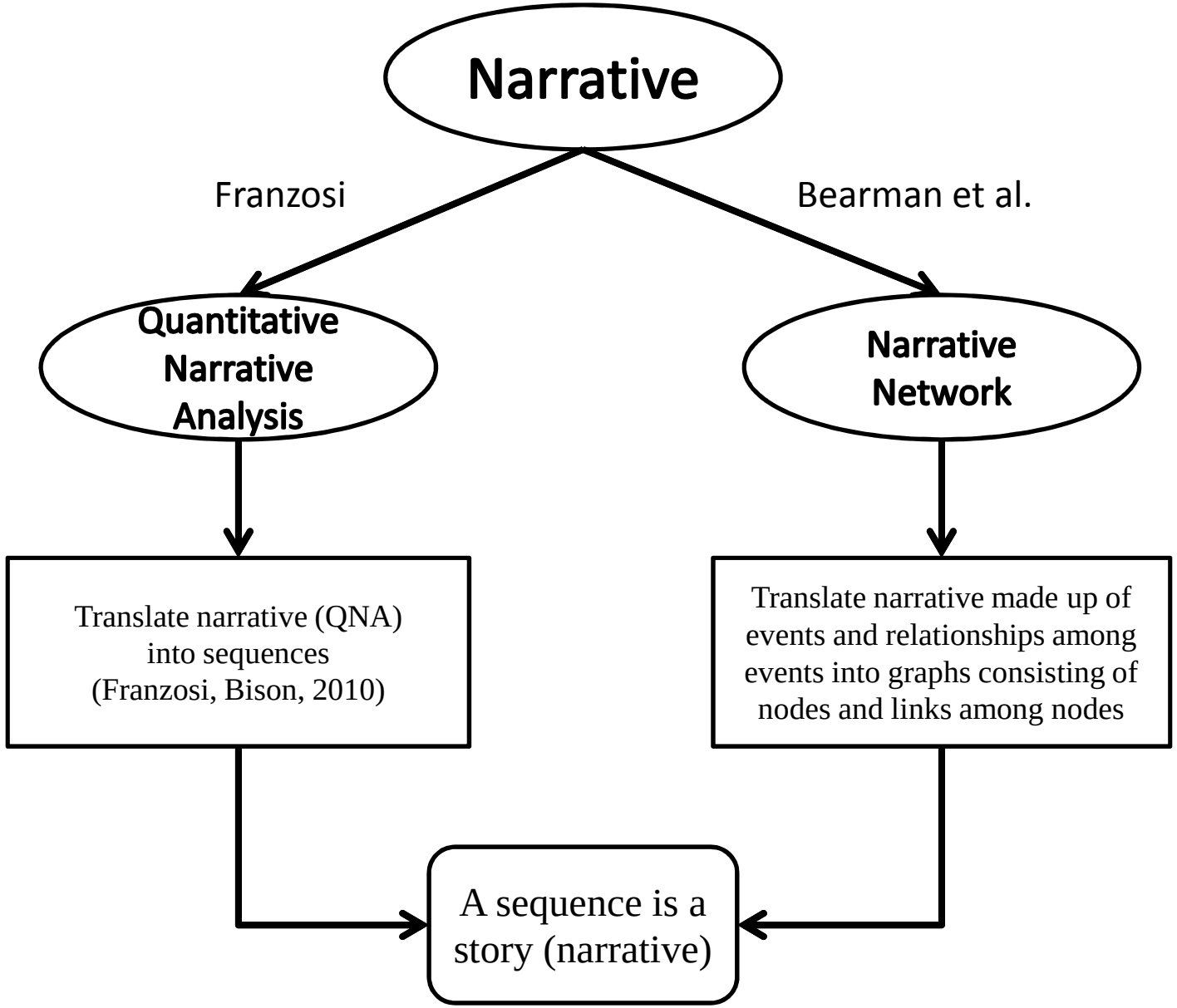
And, in the near future, also to be able **to model dynamically the processes** that have generated a specific pattern, a given sequence.

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.

A network (or a graph) is a collection of nodes (or vertices), and the connections among them are called arcs, ties, edges.





A sequence is a standardized and orderly narrative in which all the elements/events/states that compose it are temporally ordered.

From the sequences to the network

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.

{EUEE} ; {EUEU}

We can 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.

Fig. 1.0, Network plot of sequence {EUEE}

Fig. 2.0, Network plot of sequence {EUEU}

We must shift our focus from the individual sequences to what they have in common and what differentiates them.

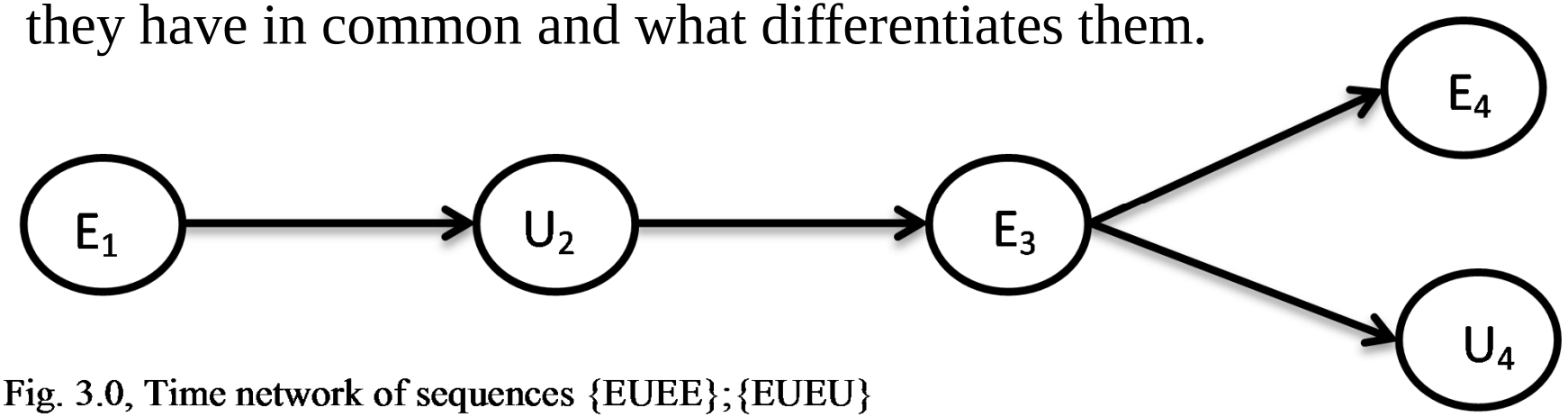


Fig. 3.0, Time network of sequences {EUEE};{EUEU}

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.

(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 frequency or 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;

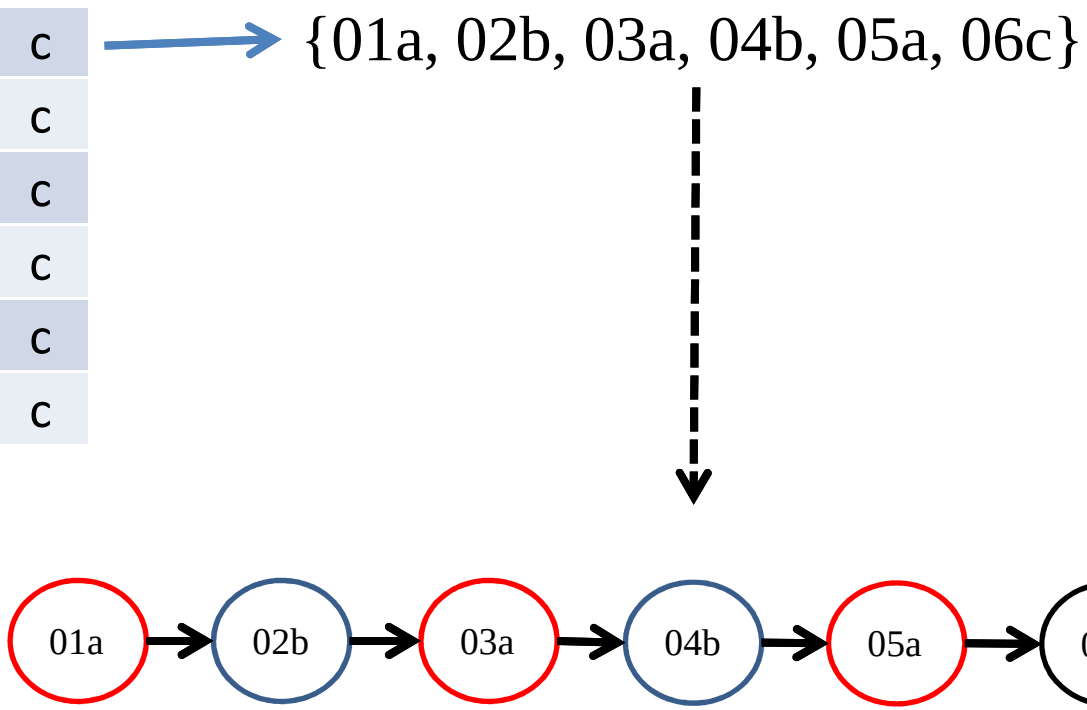
(c) Multiple nodes can be defined in the same unit of time. In the example of the two previous sequences, observed at time t4 is both an event of unemployment and one of employment.

(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.

From NetWork to NetDraw

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

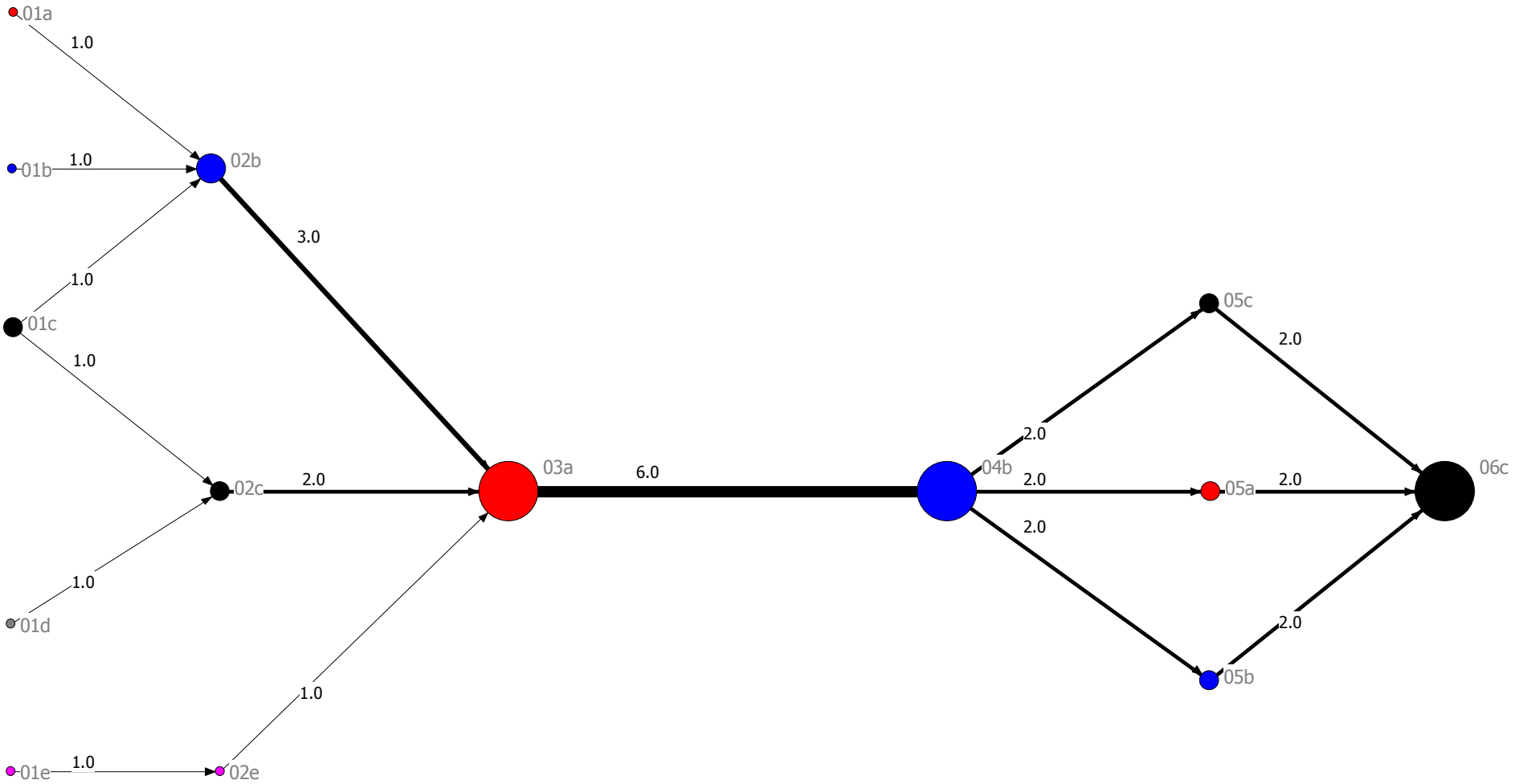
id	t1	t2	t3	t4	t5	t6
1	a	b	a	b	a	c
2	b	b	a	b	a	c
3	c	b	a	b	b	c
4	c	c	a	b	b	c
5	d	c	a	b	c	c
6	e	e	a	b	c	c



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

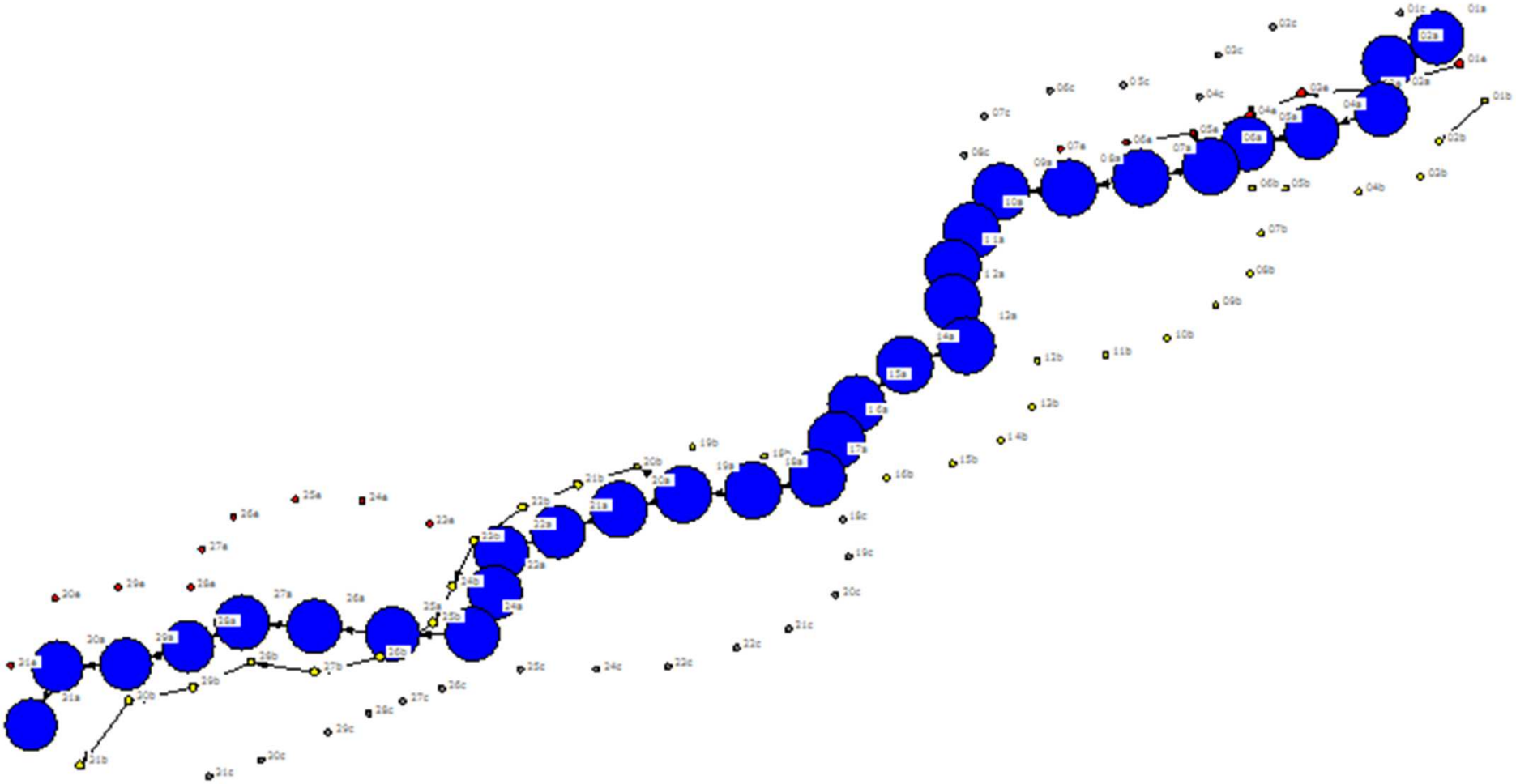
Fig. 4.0. Time sequence network of the six sequences in Table 1.0.



A real example

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

with nodes and ties weighted



Einstein was right. Time is space.

The time sequence network 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.

The problem is

what would happen if we decided to cancel time

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

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

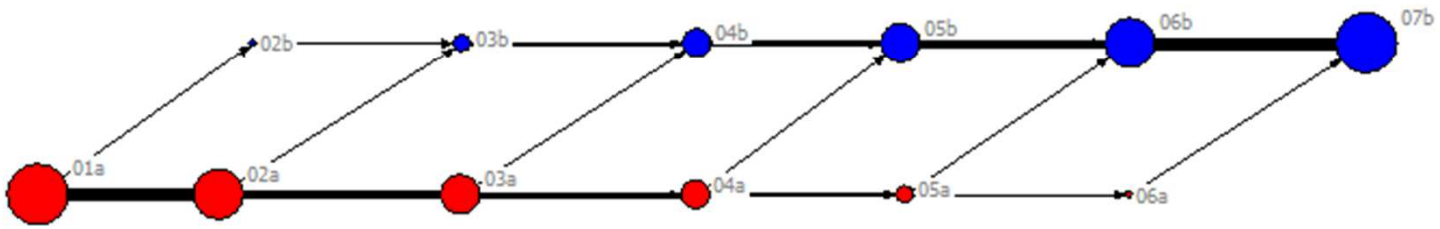


Fig.7.0. Time sequence network of the six sequences of table 4.0.

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

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

Time Position - State

{01a, 02a, 03a, 04a, 05b, 06b, 07a}

Order Position - Events

{01a, 02b, 03a}

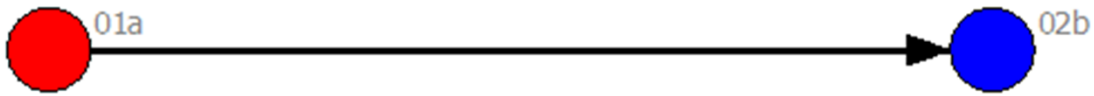
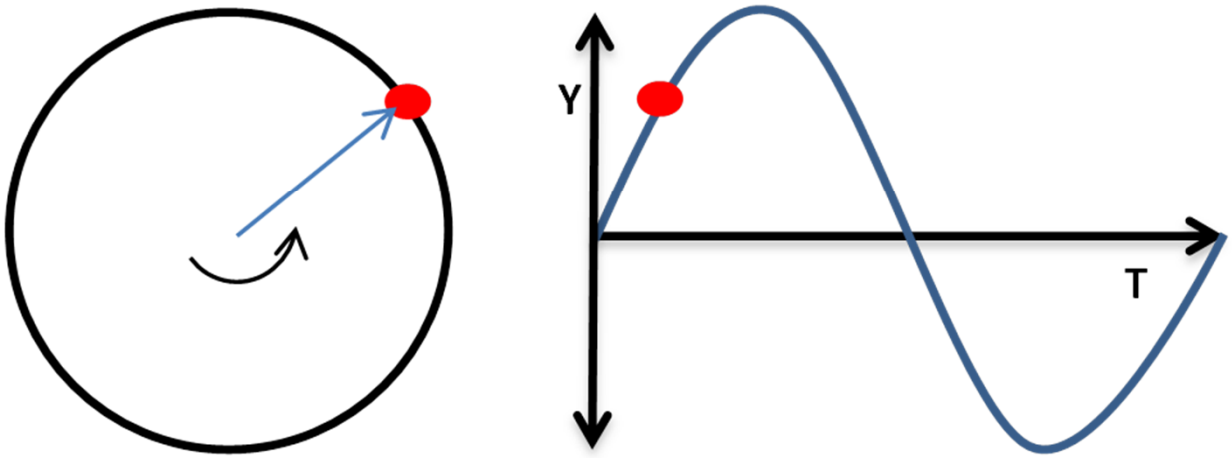
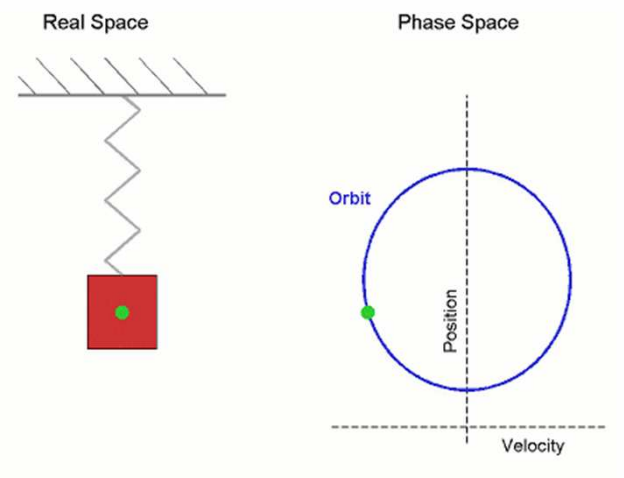


Fig. 8.0: Event sequence network of the six sequences in table 4.0.

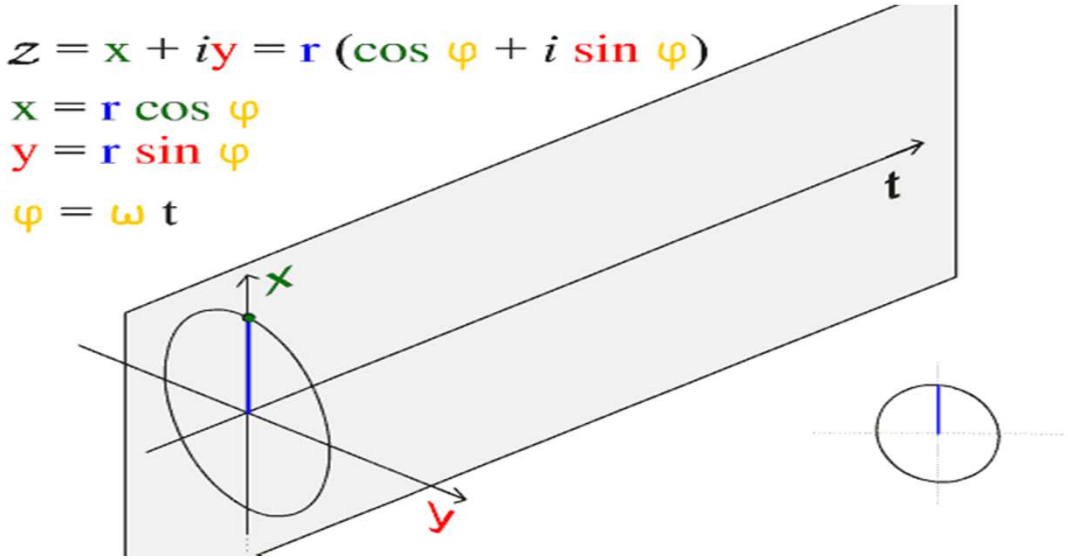
What it mean?

Figure 9.0. Simple harmonic motion:





Mazemaster, (09:11, 27 January 2009), Animation demonstrating the simple harmonic motion of a mass on a spring in both real space and phase space. Note that the phase space axes are switched from the standard convention in order to align the two diagrams. http://en.wikipedia.org/wiki/File:Simple_Harmonic_Motion_Orbit.gif



Nashev: Real sinusoid on a timebase, formed by a linear increment of complex argument in time. wikimedia.org, (09:19, 9 nov 2007), http://commons.wikimedia.org/wiki/Image:ComplexSinInATimeAxe.gif?uselang=it

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 one represents the shape of space described at the point in time; the second represents the shape of time described at the point in space.

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 (or the genetic blueprint) 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
the other describes the shape of time
together describes how the space changes over time

The careers of men and women

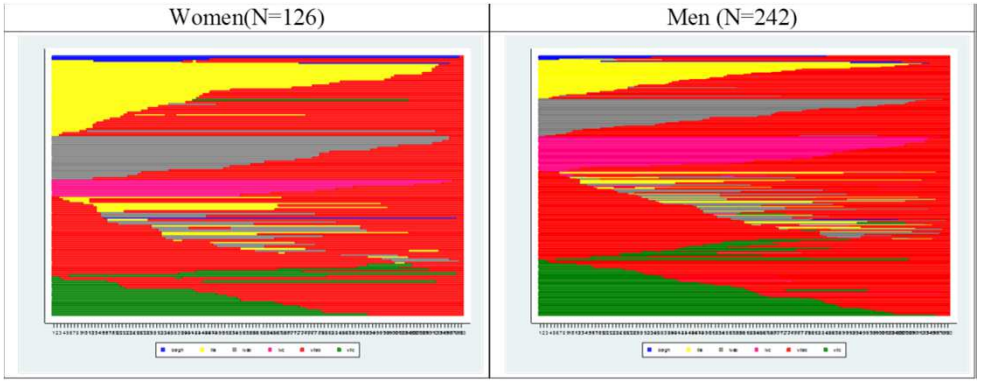


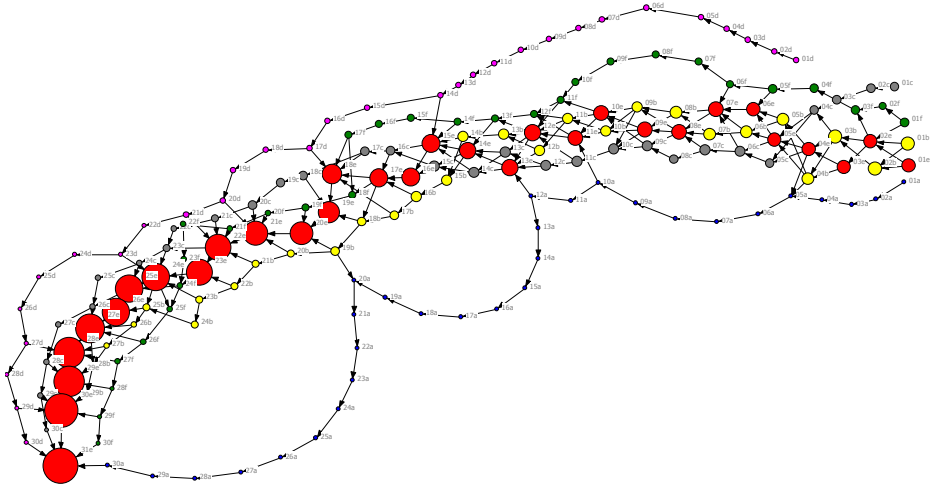
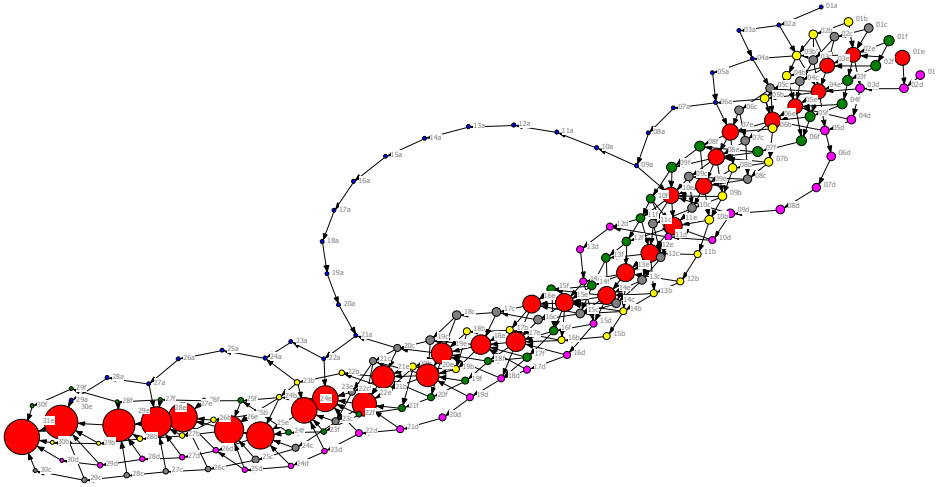
Fig.12.1 Women (N=126)	Fig.12.2 Men (N=242)
	

Fig.12.1 Women (N=126)

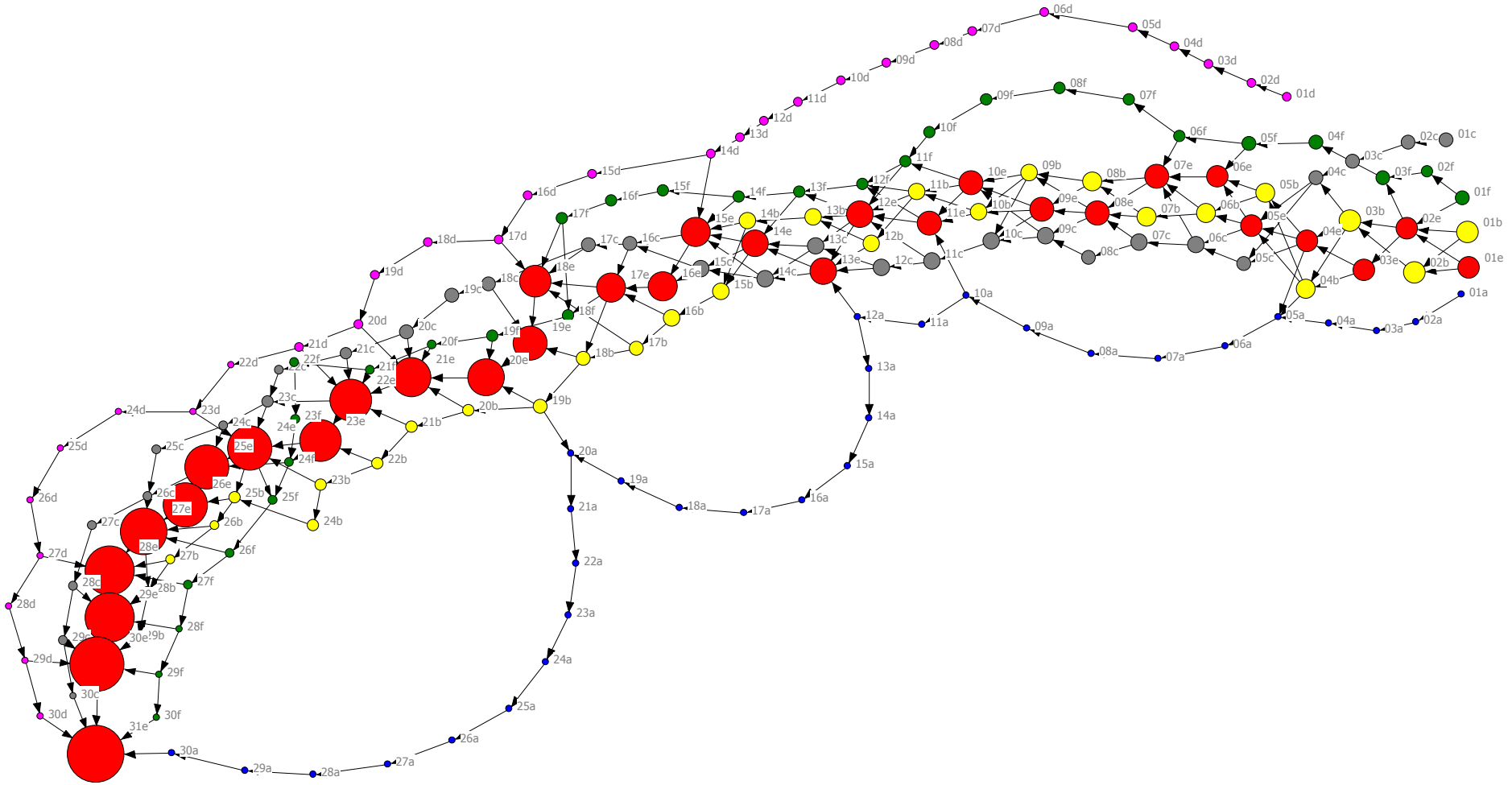
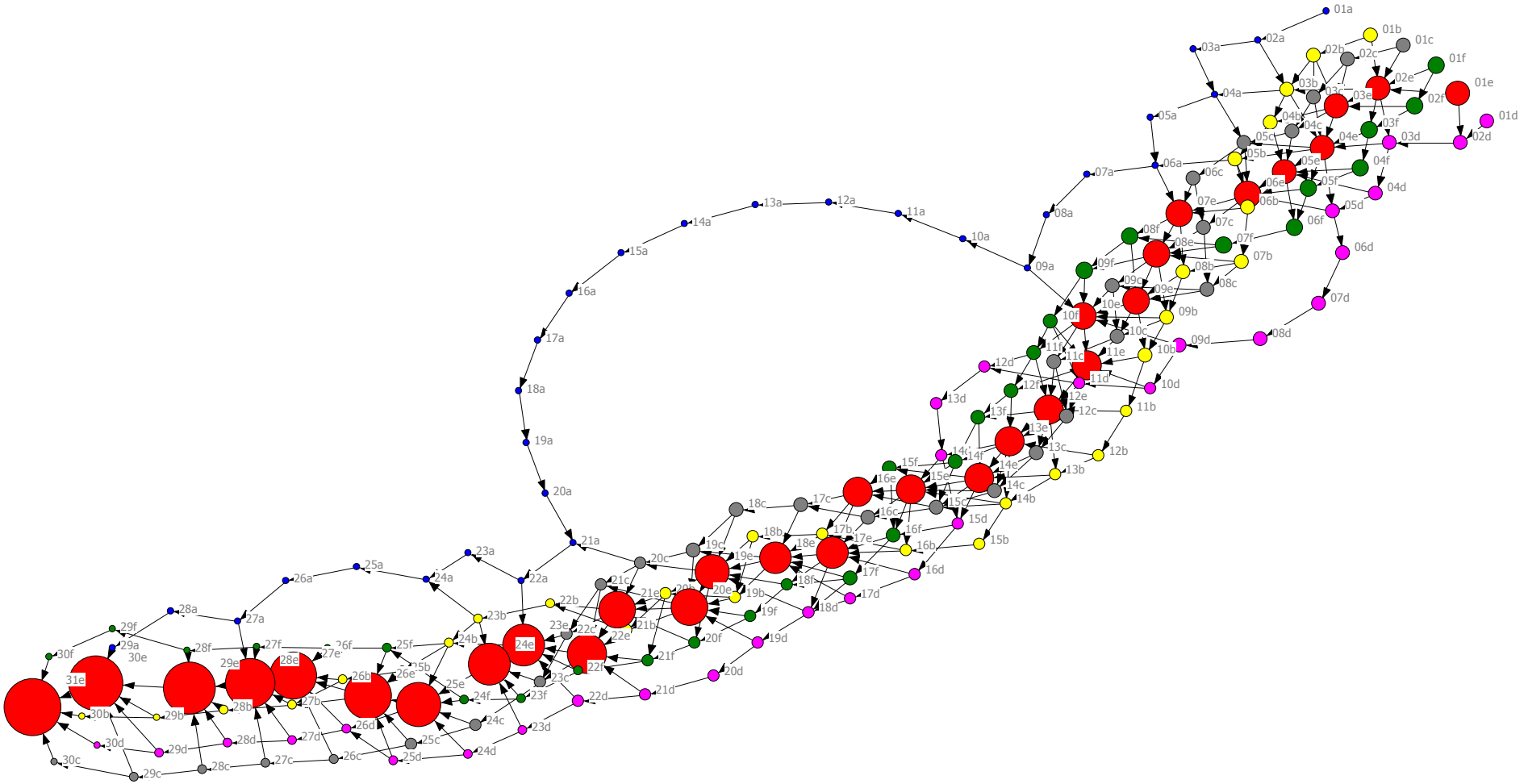
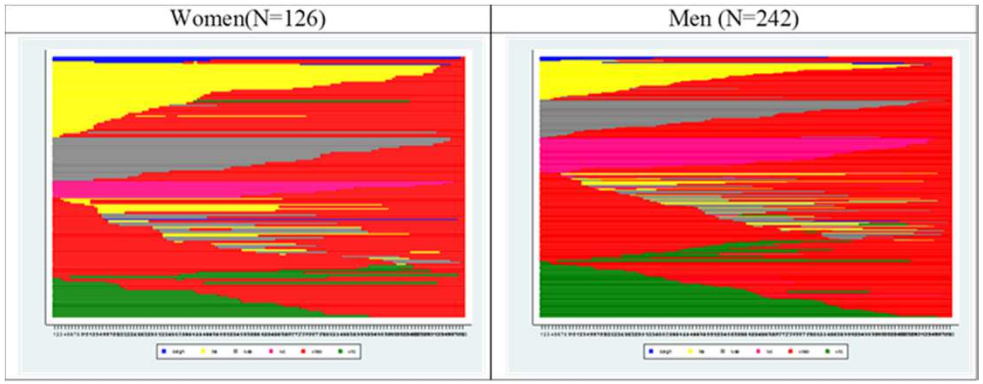
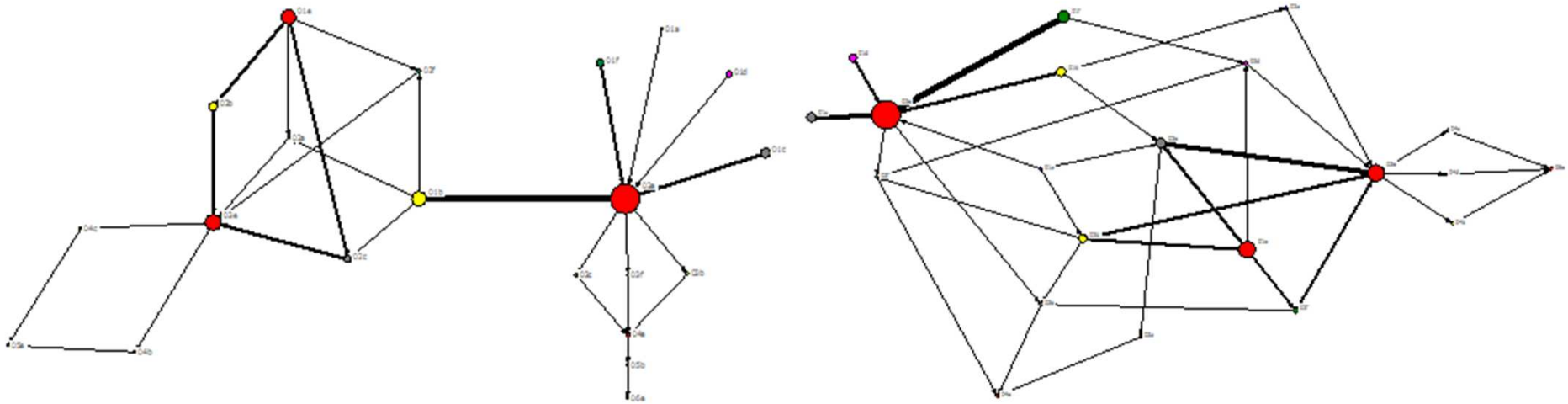


Fig.12.2 Men (N=242)







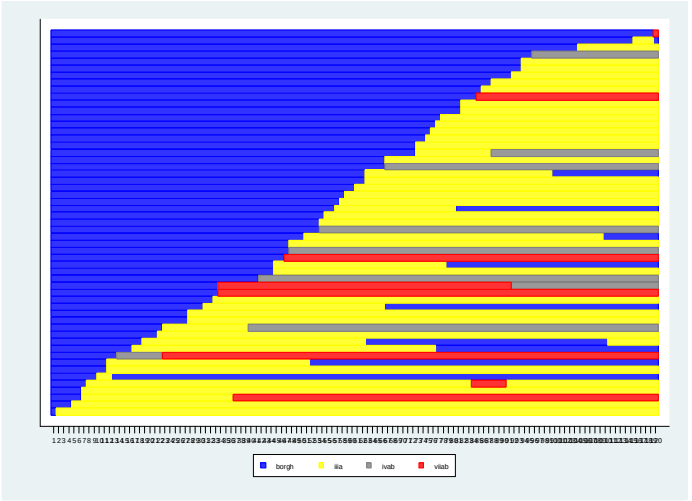
Women (N=126) Men (N=242)

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

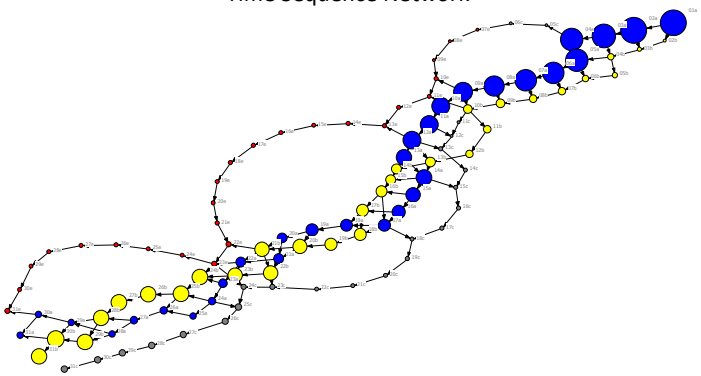
Examples with real data

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

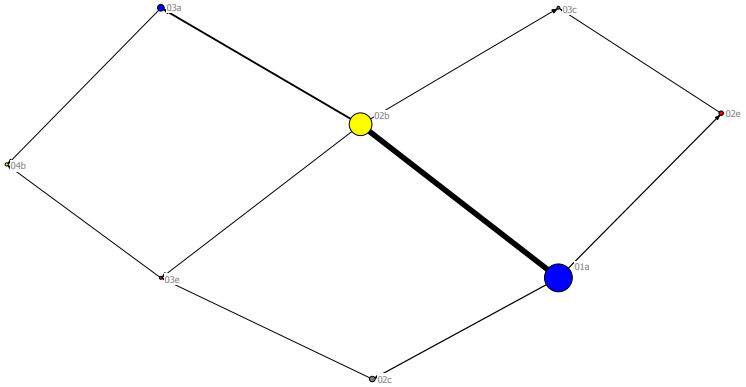
Sequence Index plot



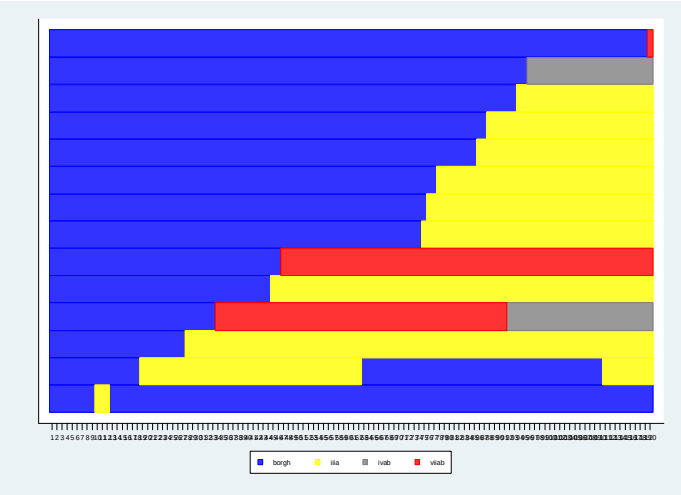
Time Sequence Network



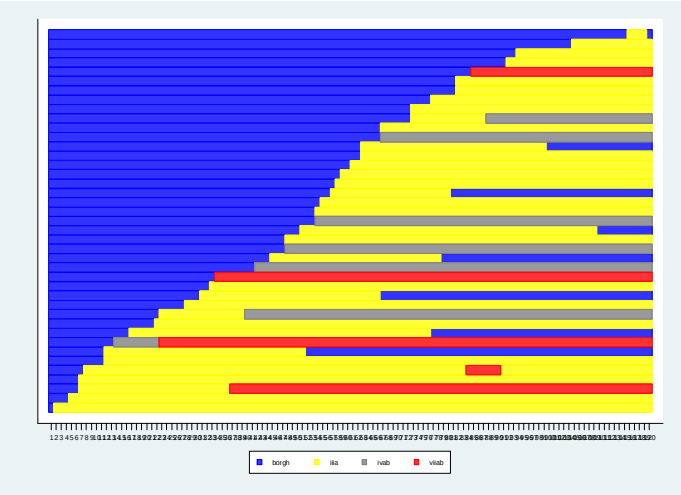
Event Sequence Network



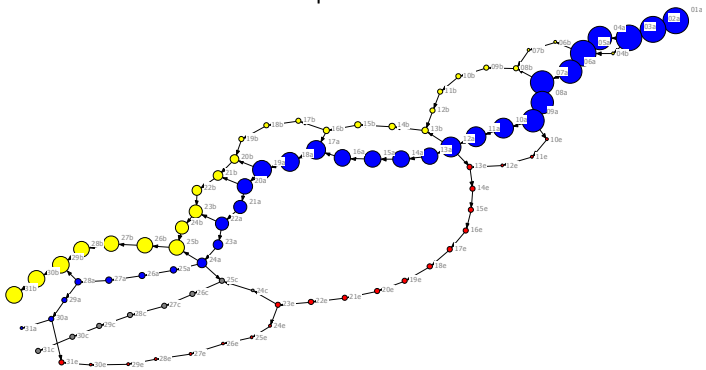
Sequence Index plot



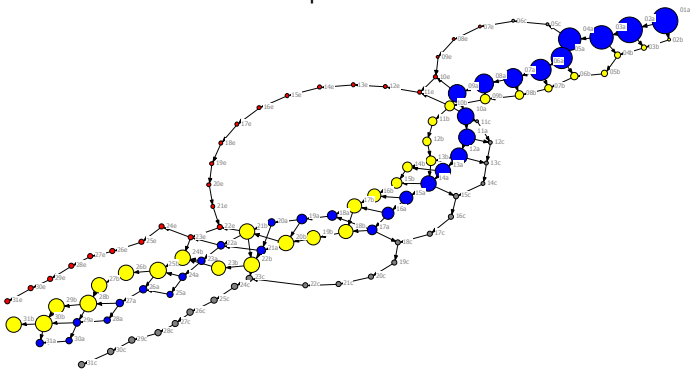
Sequence Index plot



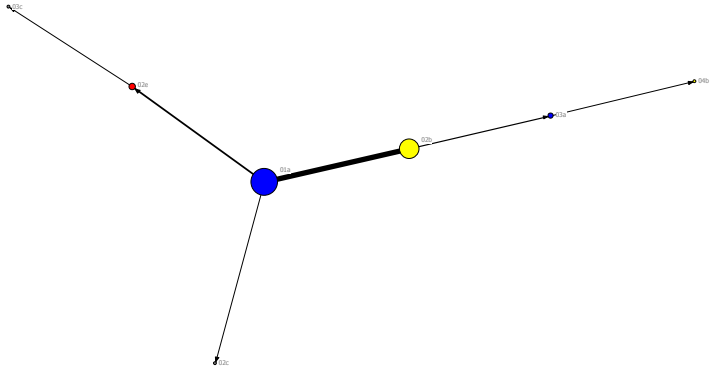
Time Sequence Network



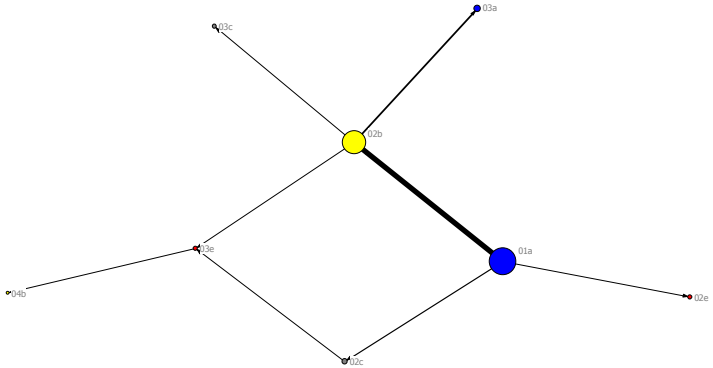
Time Sequence Network



Event Sequence Network



Event Sequence Network

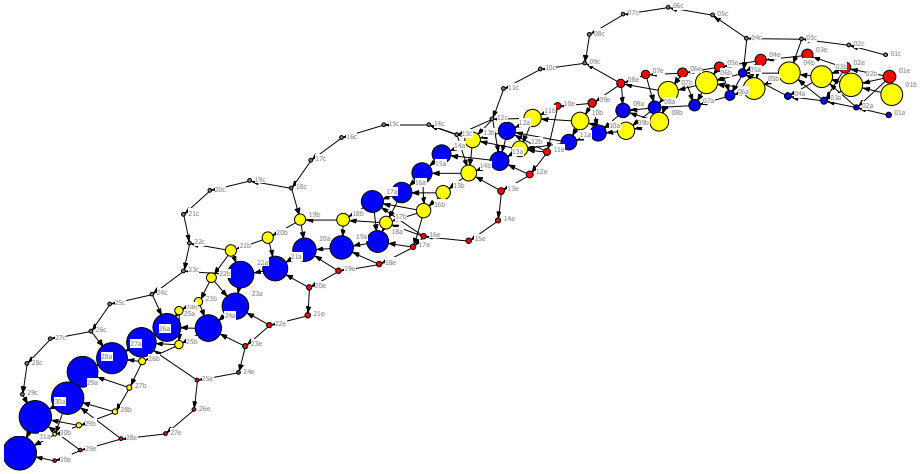


Women. (N=14)

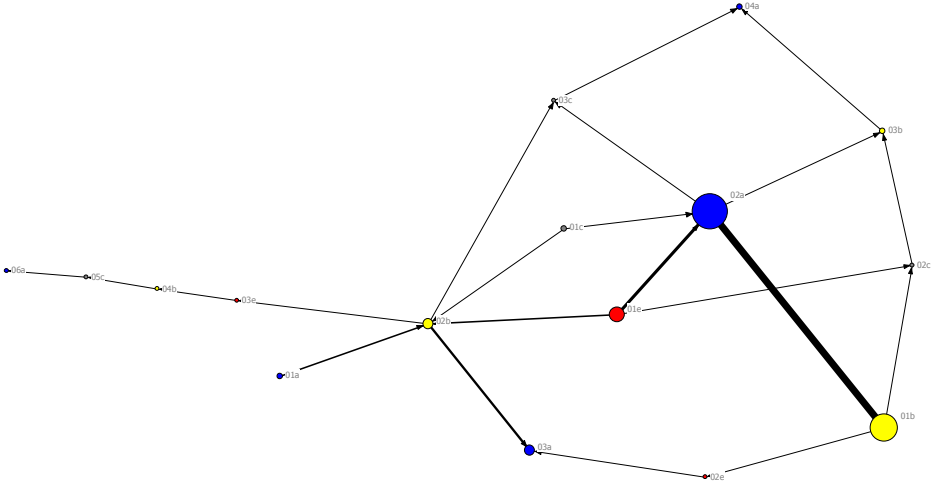
Men (N=41)

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

Time Sequence Network

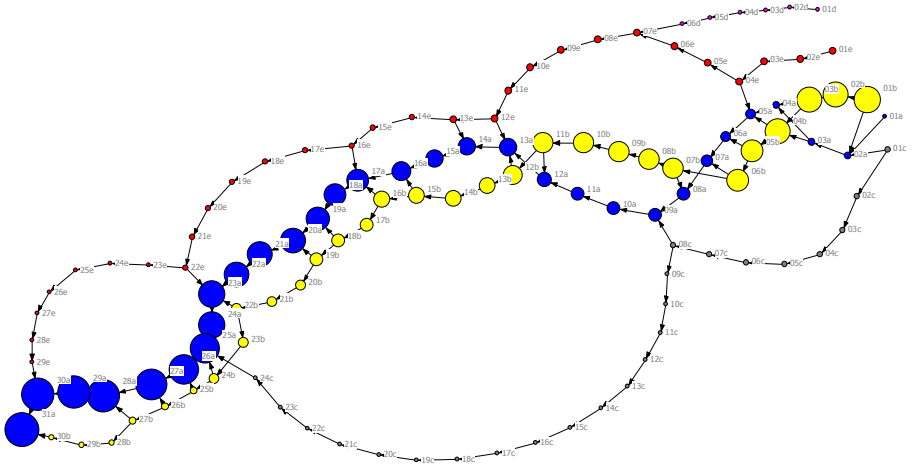


Event Sequence Network

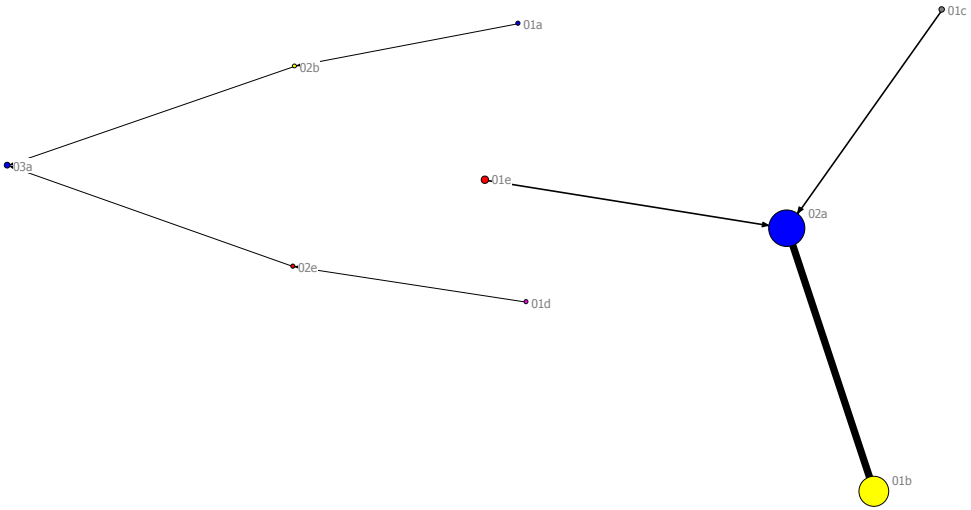


Class careers that end in class I & II. Women. (N=42)

Time Sequence Network

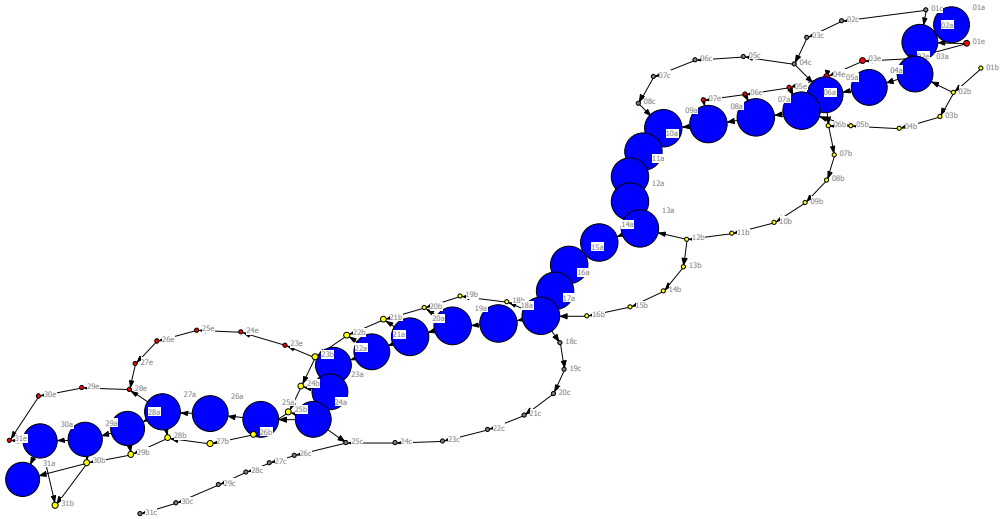


Event Sequence Network

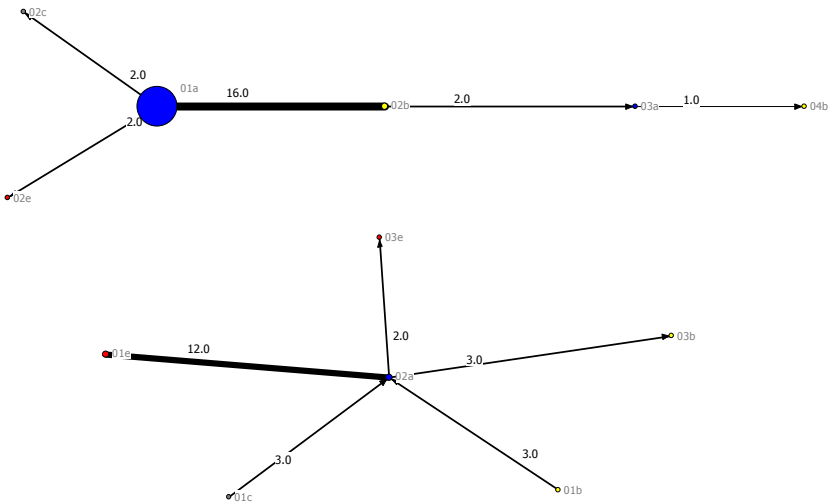


Immobility pattern I+II: Cluster a (N=247)

Time Sequence Network



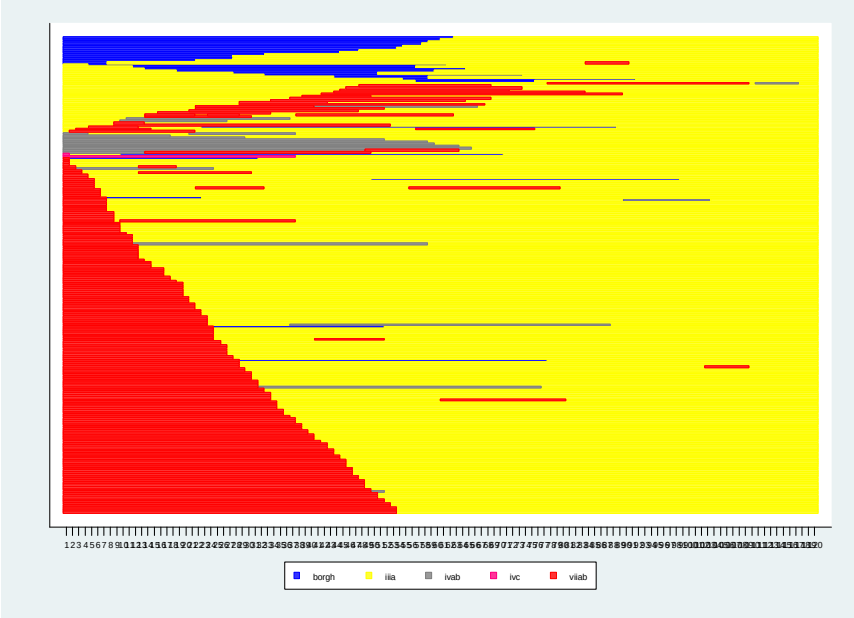
Event Sequence Network



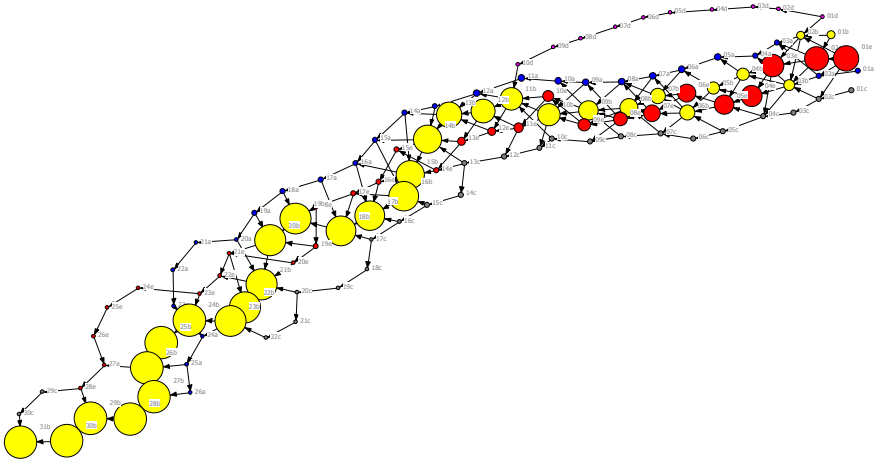
Upwards mobility patterns: IIIb+V-VI+VIIa→IIIa

(fast); Cluster l (N=229)

Sequence Index plot

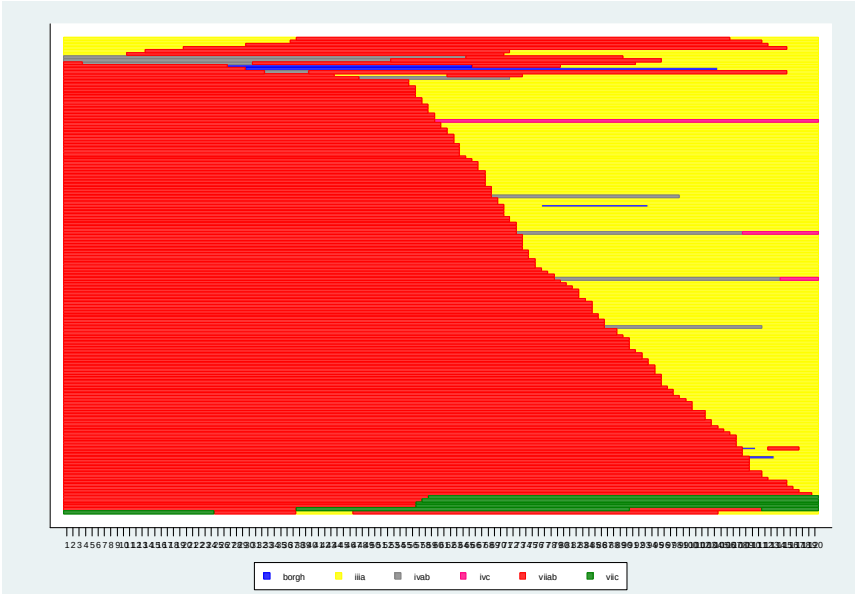


Time Sequence Network

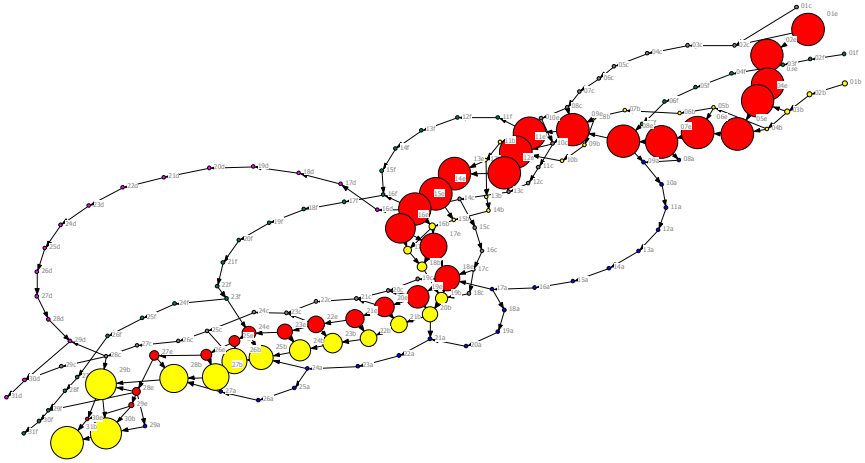


(slow); Cluster m (157)

Sequence Index plot

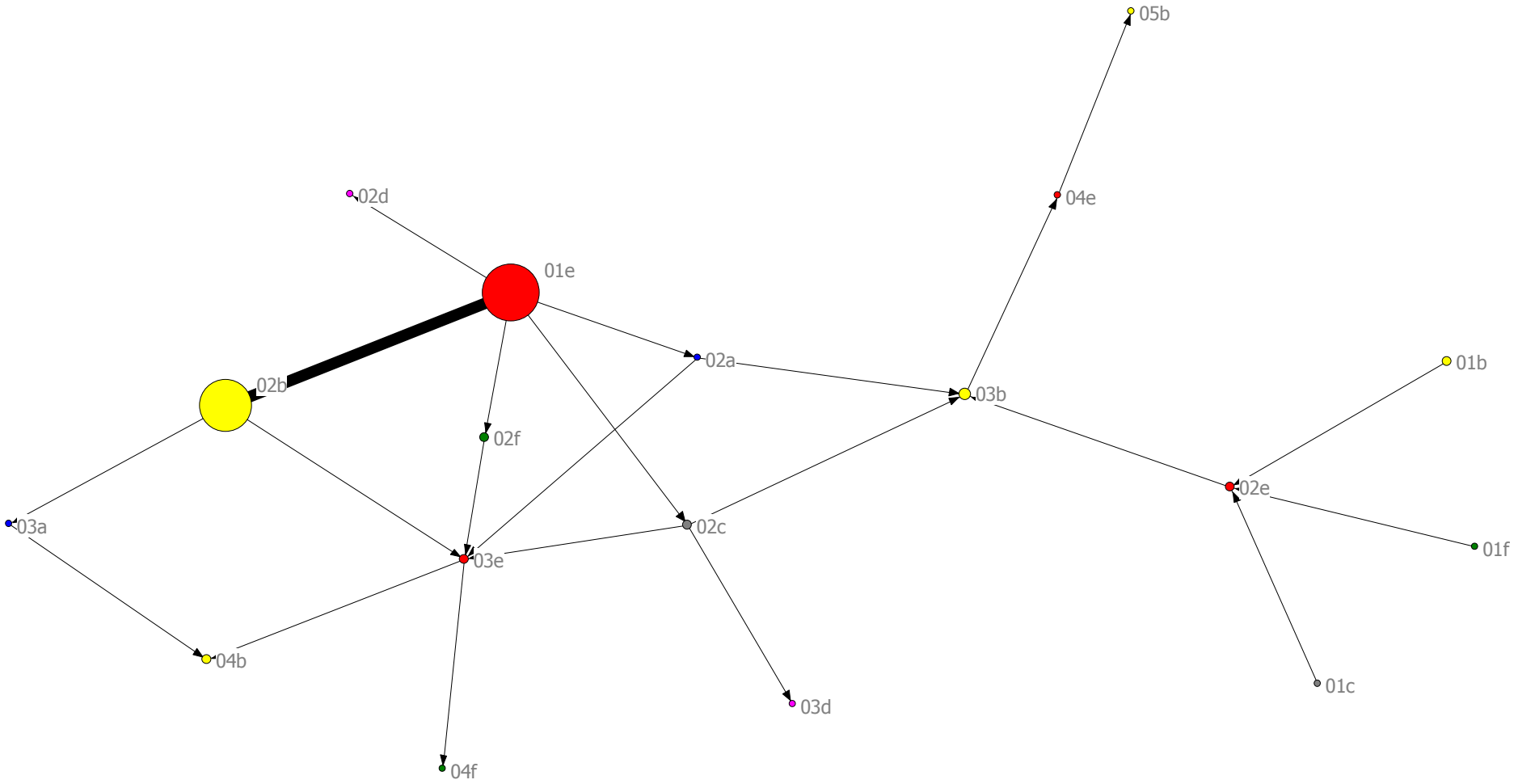


Time Sequence Network



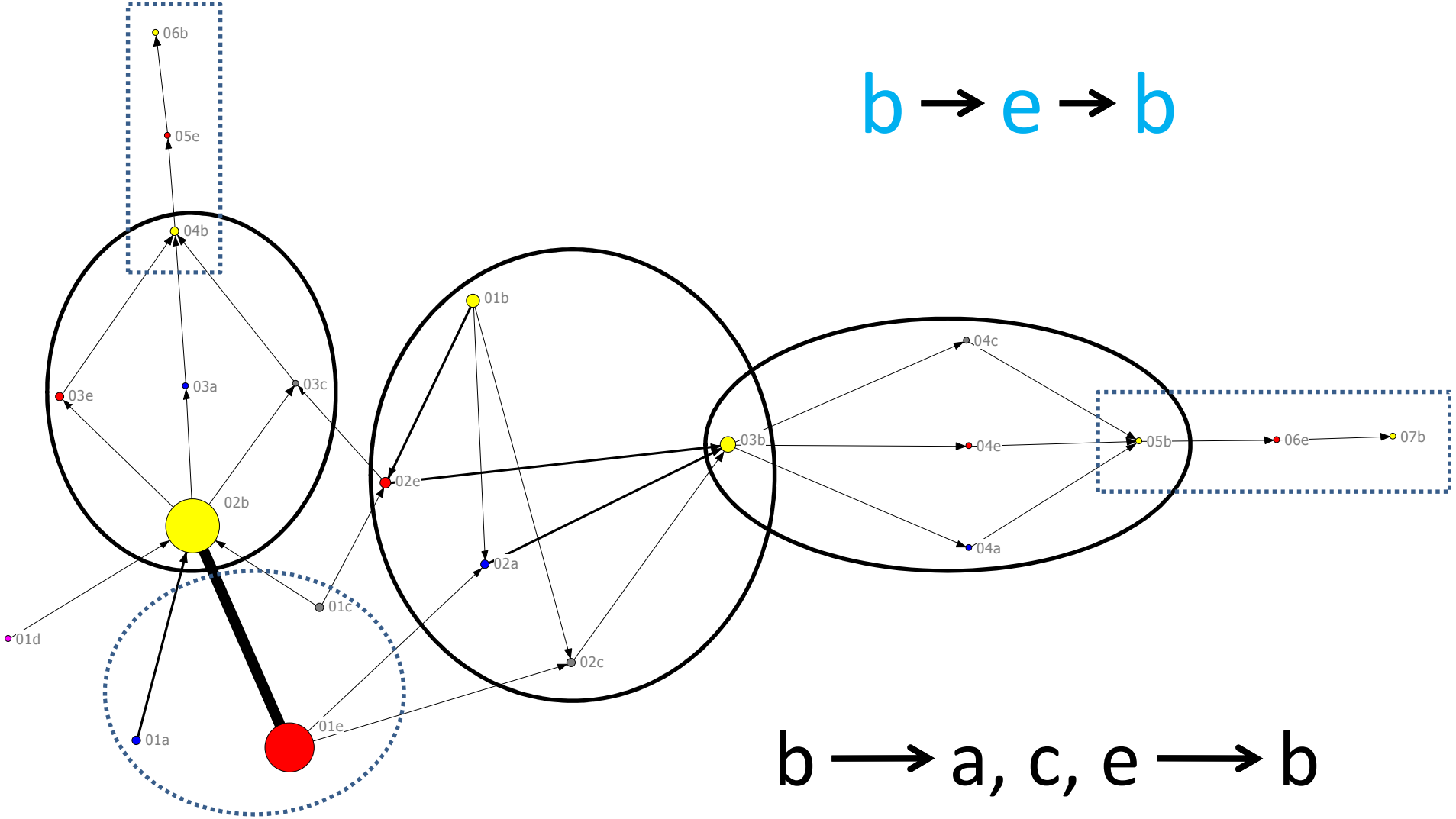
Upwards mobility patterns: IIIb+V-VI+VIIa→IIIa (slow);

Cluster m (N=157)



Upwards mobility patterns: IIIb+V-VI+VIIa→IIIa (fast);

Cluster l (N=229)



b → e → b

b ⟶ a, c, e ⟶ b

e → b ⟶ a, c, e ⟶ b → e → b

And Now?

There are many problems that remain to be solve

There are many things that remain to be done.

Problems

The use of these forms of display requires more attention from researchers.

- Extreme sensitivity of the instrument even in visualization of the more marginal patterns.

- These are complex graphics and may mislead the researcher

> *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.*

- An other methodological problems is the annulment of individual sequences

> *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. (Distribution plot, Parallel plot)*

things that remain to be done

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 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.

> *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 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.*

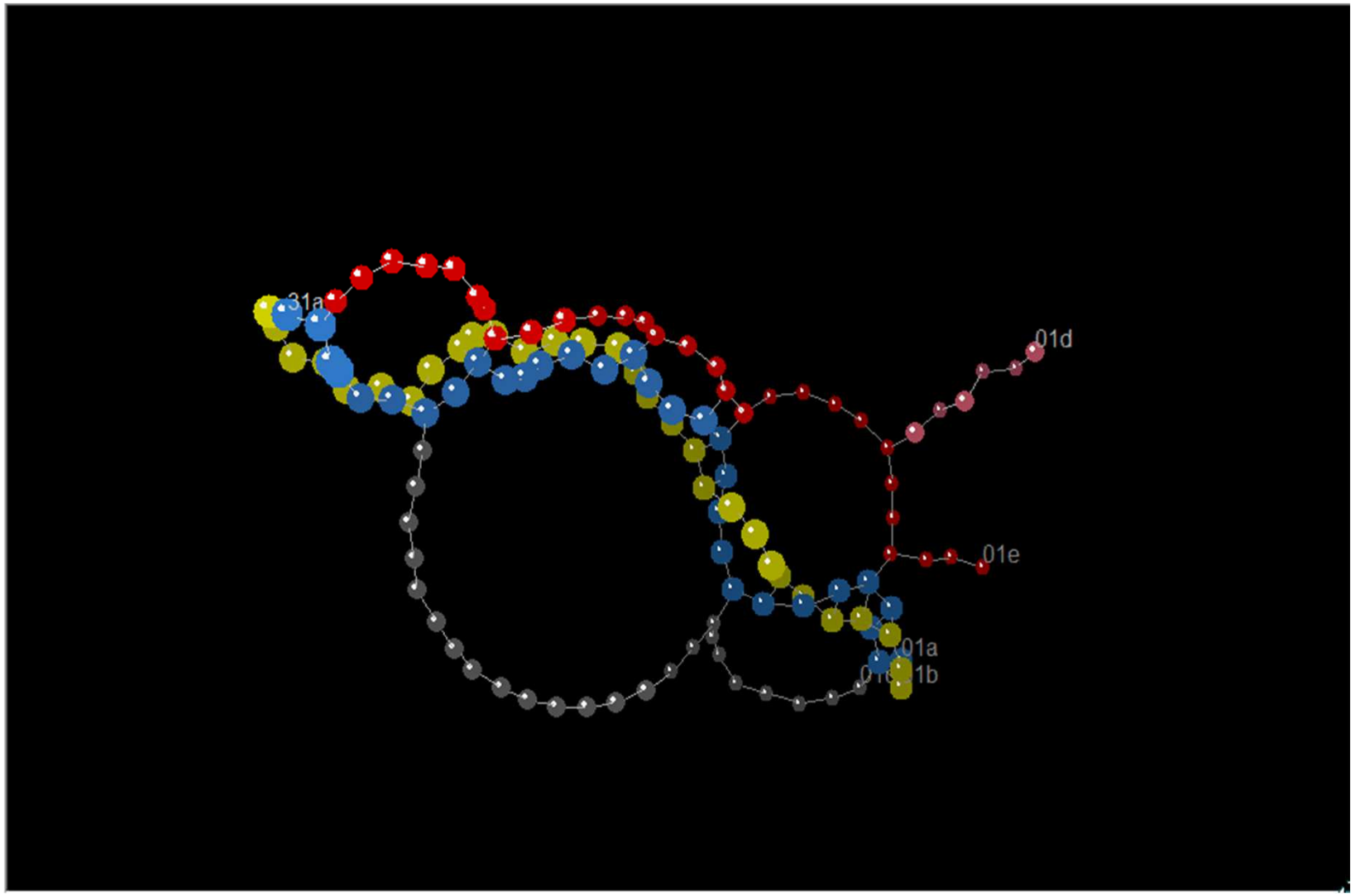
This required new tools with which to study and graphically model the evolution in time and space of our patterns.

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 that **the shift to a network perspective** could provide research tools more useful for studying sequences.

I think that when we are able to adopt these tools

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



the only limitation will be our imagination, our ability to imagine.

Thank you