FORMULAS: definition of Graph: G(V,E)…. Degree, Eig…(centrality) topologies formulas??? TESLA formula for sure 🡨 theoretical background

New proposed analysis: defined in experimental results???

https://en.wikipedia.org/wiki/Graph\_property

* **Abstract**
* **Theoretical Background (literature review + network analysis theories + tools used)**

BACKGROUND

In many problems it is helpful and even necessary to study the relationships between entities of a system since we find them interconnected forming networks in which some small changes on an element can propagate and affect the entire system. To do so we use graph, and more precisely networks, theories which represent each individuals or actors by nodes and the connection or relationships as edges between the nodes.

We find a large list of examples in which it is necessary to apply network theory on different and wide subjects as social science, finance, biology or climatology.

Studying such networks can reveal lots of information like how an event can trigger a topological change of the entire network, how entities of the network can depend on each other’s states or share similar properties and organize themselves into groups or which individuals are the main agents that lead the evolution of the network.

Classically these analysis have been carried in the simplest way assuming static (non-varying) networks, however this is usually a simplification since in most of the cases the relationships between agents in a network evolve over time as the nodes’ states change. The internet, biological processes or the economy are some examples of networks evolving over time, shrinking or creating new links between the nodes.

Although there is wide literature and methods on modelling static networks it has not been until recent years that interest has grown on studying dynamic or time-varying networks ([1], [3], [5], [6], [7], [8], [11]).

Usually, only time series measurements, such as microarray, stock price, etc., of the activity of the nodal entities, but not their linkage status, are available. The goal is to recover the latent time-varying networks with temporal resolution up to every single time point based on time series measurements.

To be able to understand the current state of the research on the field and the strengths, weakness and challenges of the techniques proposed this work is going to analyze several algorithms, focusing mainly on TESLA (from the acronym TESLLOR, which stands for temporally smoothed *l*1*-*regularized logistic regression), built to recover the structure of time-varying networks over a fixed set of nodes from their time series nodal attributes [1].

# network main properties

Depending on the problem under consideration, it may be convenient to study the network from different points of view, focusing on relationships between the entities (edge-centric), one or more entities (vertex-centric) or the topological properties of the network as a whole (graph-centric) [5].

When studying edge-centric networks some important measures include transitivity (similar to the mathematical property: if *u* connected to *v* and *v* connected to *w*, then *u* connected to *w*) and reciprocity (when we find loops in a network with 2 edges length, so pairs of nodes that have edges running in both directions).

Meanwhile in vertex-centric networks centrality measures are used to quantify the relevance of the different nodes in a network. The most basic example is the degree of centrality of the nodes: the number of edges attached to them. Nodes with high degree are called hubs and obviously play an important role in the system.

Regarding graph-centric approach we are able to analyze the general shape by studying the clusters in which the vertices organize themselves, many techniques like the well-known hierarchical clustering are useful in this task showing important properties of the network like its tendency to follow an assortative or dissortative mixing pattern.

All these properties [9] give us important information to understand the behavior of the system.

# dynamic networks

Latest years the science of networks developed three important aspects that define this field nowadays: it is used to model real-world problems, it frequently assumes the networks are not static and it aims to understand networks not just as topological objects but also as the framework upon distributed dynamical systems are built [3].

Then studying networks as dynamical systems is the best way to understand evolving underlying processes in complex systems that can trigger critic events reshaping the network and changing its properties, one example of this is how a disease spreads: the structure of a network through which a contagious agent is transmitted can have a dramatic impact on outcomes at the level of entire populations [3].

# learning networks

Looking at their topology we find that networks are organized in two main classes: scale-free networks and random graphs.

Scale-free network term was coined to describe the networks that show a power law degree distribution (distribution of the connections over the whole graph) after Barabasi mapped a portion of the World Wide Web [2] finding that it followed this distribution of the links and that some nodes had many more links than others (hubs). Then the most notable characteristic in scale-free networks is that it is relatively common to find vertices with a much higher degree (links) than the average. They also have other important features like robustness due to its hierarchical organization or clustering coefficient that may help to understand networks following this distribution.

A random graph is obtained by starting with a set of isolated vertices and adding successive edges between them at random. The aim of the study in this field is to determine at what stage a particular property of the graph is likely to arise. Different random graph models produce different [probability distributions](https://en.wikipedia.org/wiki/Probability_distribution) on graphs [4].

## Exponential Random Graph Model

Based on this last one, the exponential random graph model (ERGM) has been extensively used to model static networks and also used as a base for two of the first algorithms proposed to study dynamic networks, the temporal exponential random graph model (tERGM) and the hidden tERGM for modelling a sequence of node attribute [7]. When tERGM assume that the sequence of networks is available, htERGM explores the possible dependencies of unobserved rewiring networks and leads to the algorithm that can reconstruct such networks from a snapshots’ sequence of nodal attributes.

Although this algorithms overcame approaches that recover single time-invariant networks they are not the most useful ones since they depend on unobserved network variables being unable to compute likelihood ratios, needing inference algorithms specifically for each problem.

## Hidden Markov Dynamic Bayesian Network

Another classical approach are the Bayesian networks, widely used to describe biological systems, risk analysis and even financial networks.

Bayesian networks are graphs whose edges represent conditional dependencies and nodes are random variables in a Bayesian sense (observable quantities, latent variables, unknown parameters or hypotheses) and are associated with probability functions that takes as input the node’s parent variables and gives the probability of the variable represented by the node.

Its standard assumption is ‘stationarity’, and therefore, several research efforts have been recently proposed to relax this restriction. However, those methods suffer from three challenges: long running time, low accuracy and reliance on parameter settings. [11] propose a non-stationary DBN model by extending each hidden node of Hidden Markov Model into a DBN (called HMDBN), which properly handles the underlying time-evolving networks resulting in a promising experimental evaluation of the method, demonstrating more stably high prediction accuracy and significantly improved computation efficiency (even with no prior knowledge and parameter settings) on both synthetic and real biological data.

Although this probabilistic model is more complex than TESLA algorithm it could be a good method to implement in order to compare results and performance.

## Bayesian non-parametric model

In this line there was proposed a Bayesian non-parametric model including time-varying predictors in dynamic network inference precisely for financial studies [6].

This model computes edge specific predictors where the link probabilities () are estimated via a logistic regression, with a baseline process quantifying the overall propensity to form links in the network across time, are vectors containing the latent coordinates favoring a higher link probability when units *i* and *j* have latent coordinates in the same direction and is a P-dimensional vector of time-varying edge-specific predictors for units *i* and *j* at time t and are the corresponding dynamic coefficients. This allows the proximity between units *i* and *j* at time *t* to depend on predictors in a manner that varies smoothly with time.

The main issue regarding this method is that it assumes time-constant smoothness while in finance and other network applications we expect smoothness to vary over time, however it had some promising results when analyzing the global stock market network in the 2008 crisis and this makes it a good candidate to test in this project.

## Temporally smoothed l1-regularized logistic regression

TESLA represents an extension of the lasso-style sparse structure recovery technique and is based on a key assumption that temporally adjacent networks are likely not to be dramatically different from each other in topology and therefore are more likely to share common edges than temporally distant networks.

Building on the *l*1-regularized logistic regression algorithm for estimating single sparse networks [10] it was developed a regression regularization scheme that connects multiple time-specific network inference functions via a first-order edge smoothness function that encourages edge retention between time-adjacent networks. An important property of this idea is that it fully integrates all available samples of the entire time series in a single inference procedure, what means an advantage in contrast of htERGM algorithm.

TESLA estimates , the correlation (or dependency strength) matrix between the nodes using a time-series of the observed stated of the nodes **xt**.

where,

Where the graph structure is given by the locations of the nonzero elements of the parameter vectors . Components of the vectors are indexed by distinct pairs of nodes and a component *j* of the vector is nonzero if and only if the corresponding edge (*i, j*) ∈ Eτ. denotes the observed states of all nodes but node *i* and is the log conditional likelihood of state under a logistic regression model.

First term of the algorithm are p logistic regressions, one for each node with respect the rest of them ( are the states of all nodes but node I in the dth sample in time epoch t).

The second term is a lasso regularization that introduces sparsity and consistency in neighborhood selection, as showed in [10], converging to the true graph structure. Also a sparse graph effectively limits the degree of freedom of the model, which makes structure recovery possible given a small sample size [8].

Since in this problem we are estimating dynamic networks (so more than one single graph), and, as assumed before, temporary adjacent networks are going to be similar one to each other, we need to introduce the third term in the equation that penalizes the discrepancy between time-adjacent parameters .

Summarizing, the first term is the main one while the last two terms are the penalties () that are introduced to enforce sparsity and smoothness.

## TESLA updated

It is important to note that in this specific method we are assuming that are a piecewise constant function with abrupt changes in parameters (jumps).

However in some cases the changes in parameters are continuous, so further work [8] was carried out to adapt this algorithm including a weighting term that turns in continuous functions instead of the penalty in the structural changes () that assumes small “jumps” between time-adjacent values.

Where the weights are defined by

And is a symmetric nonnegative kernel function.

# financial application

When studying dynamic networks we aim to solve real world problems, so it is usual to look at meta-networks, this is multi-link (many types of links), multi-mode (many types of nodes) and multi-level (nodes can represent subnetworks) networks.

This added to the fact that not many work has been done to recover dynamic networks on finance and the promising algorithms reviewed on this paper sum up enough motivations to build an interesting topic: testing the performance of these different methods in the financial area to find the model that best defines it and extending the previous work done in this subject (analyze the behavior of financial networks against critic events [6]) by looking at meta-networks instead of simple networks.

Define networks:::: define directed graph and note that this is the case we are using here (modelling this way to take into account 2 different kinds of itneractions: InDeg->popularity, OutDeg->gregarious)

Network dynamics: – The network topology changes over times – Nodes and/or edges may come and go – Captures faults and reliability issue

These dynamic networks suffer an evolution ruled by the interactions between the individuals. One of the main assumptions when modelling these networks is that they tend to reach the equilibrium state, the states TESLA recovers.

A way to find an equilibrium state and measure how far the network is, so how unbalance the network is???

By duration networks can be – Transient: The dynamics occur for a short period, after which the system is static for an extended time period – Continuous: Changes are constantly occurring and the system has to constantly adapt to them

After all: It makes sense to model the financial market as a dynamic network: Network since its elements are linked and their behaviours can affect the whole system and dynamic to model the highly changing character of the stock market, that difficulties the ability to model it as a static system. In addition modelling a system as a evolving network gives the potential to study patterns or events’ effects that affect the network.

**NETWORK ANALYSIS**

**Intro**

When analysing a network we have to focus on different levels, going from the overall network as whole (in order to understand its distribution and measure the cohesion of the network), and later focusing on the individuals by measuring their centralities (to identify the central nodes of the system that support the network and study movements and interactions) cross boundary analysis??

**Net topology**

To study the topology of the network first we need to find a suitable layout to build the network.

Many options are available, after testing and studying the designed layouts existing in R there were chosen some of them in order to try different approaches:

- Frutcherman-Reingold layout: based on a directed force algorithm which takes the nodes as charged particles that tend to repel while the edges are treated as springs with an [attraction force or constant???] equals to their weights (TESLA output). The algorithm iterates until it converges to a minimum energy state (equilibrium). Force directed algorithms seem an ideal choice in this case since heavily connected individuals will tend to group while keeping unrelated nodes separated what is a very intuitive way to understand their interactions. The only counterpart of this family of algorithms is that the minimum energy state which it converges is a local minimum, this means that the results don’t need to be the best fit. This equilibrium state will depend on the initial coordinates assigned to the nodes before starting to iterate, A combined application of different algorithms is helpful to solve this problem \*\*\*[Collberg, Christian; Kobourov, Stephen; Nagra, Jasvir; Pitts, Jacob; Wampler, Kevin (2003), "A System for Graph-based Visualization of the Evolution of Software", [*Proceedings of the 2003 ACM Symposium on Software Visualization (SoftVis '03)*](http://www.researchgate.net/publication/2851716_A_System_for_Graph-Based_Visualization_of_the_Evolution_of_Software/file/32bfe510fcc3a2ac65.pdf) (PDF), New York, NY, USA: ACM, pp. 77–86; figures on p. 212, [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1145/774833.774844](https://dx.doi.org/10.1145%2F774833.774844), [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [1-58113-642-0](https://en.wikipedia.org/wiki/Special:BookSources/1-58113-642-0), ]\*\*\* For example, using the Kamada–Kawai algorithm to quickly generate a reasonable initial layout and then the Fruchterman–Reingold algorithm[]](https://en.wikipedia.org/wiki/Force-directed_graph_drawing#cite_note-fr91-12) to improve the placement of neighbouring nodes.

- Kawai-Kamada layout: It is also based on a directed force algorithm. Edges:springs???

- other force directed algorithms????

- assigning discrete weights to different ranges of the TESLA weights and plot fr/kk: linspace(TESLAw(min),TESLAw(max),20), or studying their distribution (multimodal?: nModes=nDiscreteWeights…) Ej: in AnalysisPlots.pdf, slide t=1 we can observe the weights follow a multimodal Gaussian distribution, assigning the means of the weights to the nearby values ((kmeans???))

- Multidimensional scaling layout: This performs the classic MDS algorithm that uses the chosen distance metric to compute nodes similarities and then plot them in a 2D space keeping these distances and showing the edges.

This way one can directly identify similar stocks and compare with the results given by the force directed algorithms. In addition it has the potential to be a useful layout to study flows and contagion effects.

- Random/Circular/By Sectors… FIXED OVER TIME: (define the used algorithm)

Keeping the nodes fixed over time is the best way to focus on the evolution of the relationships in the system. Also this one seems to be a good candidate to study flows and compare with MDS.

- clustered by sectors (grid and starting coord for F-D

- using MDS as starting coord for force directed??

<https://en.wikipedia.org/wiki/Graph_drawing>

<https://en.wikipedia.org/wiki/Preferential_attachment>

**Measures of connection**

* Density • Shows overall level of connection within a network. • We can also look at ties within and between groups. ƒ Distance • Shows average distance for people to get to all other people. • Shorter distances mean faster, more certain, more accurate transmission / sharing.
* Shortest path: In [graph theory](https://en.wikipedia.org/wiki/Graph_theory), the **shortest path problem** is the problem of finding a [path](https://en.wikipedia.org/wiki/Path_(graph_theory)) between two [vertices](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) (or nodes) in a [graph](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)) such that the sum of the [weights](https://en.wikipedia.org/wiki/Glossary_of_graph_theory#Weighted_graphs_and_networks) of its constituent edges is minimized.
* Different algorithms have been introduced to solve this problem, considering that some of the weights of the edges may be negative the suitable one of this case is the [Bellman–Ford algorithm](https://en.wikipedia.org/wiki/Bellman%E2%80%93Ford_algorithm). (ref, formula etc)
* This can be useful to study the intra and interconnectivity of given sectors (compare with corr?) by averaging(+-sd) the shortest paths between same sector and between different sectors!

Connectivity: In [mathematics](https://en.wikipedia.org/wiki/Mathematics) and [computer science](https://en.wikipedia.org/wiki/Computer_science), **connectivity** is one of the basic concepts of [graph theory](https://en.wikipedia.org/wiki/Graph_theory): it asks for the minimum number of elements (nodes or edges) that need to be removed to disconnect the remaining nodes from each other \*\*[Diestel, R., [Graph Theory, Electronic Edition](http://diestel-graph-theory.com/GrTh.html), 2005, p 12]\*\*

 It is closely related to the theory of [network flow](https://en.wikipedia.org/wiki/Flow_network) problems. The connectivity of a graph is an important measure of its resilience as a network.

In an [undirected graph](https://en.wikipedia.org/wiki/Undirected_graph) *G*, two [*vertices*](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) *u* and *v* are called connected if *G* contains a [path](https://en.wikipedia.org/wiki/Path_(graph_theory)) from *u* to *v*. Otherwise, they are called *disconnected.*

A [directed graph](https://en.wikipedia.org/wiki/Directed_graph) is called *weakly connected* if replacing all of its directed edges with undirected edges produces a connected (undirected) graph. It is connected if it contains a directed path from *x* to *y* or a directed path from *y* to *x* for every pair of vertices *x, y*. It is *strongly connected* or *strong* if it contains a directed path from *u* to *v* and a directed path from *v* to *u* for every pair of vertices *u, v*. The [*strong components*](https://en.wikipedia.org/wiki/Strongly_connected_component) are the maximal strongly connected subgraphs> map the network as societies of strong connected components??🡪 <https://en.wikipedia.org/wiki/Strongly_connected_component> ::: modular decomposition?? (no creo)

[Bang-Jensen & Gutin (2000)](https://en.wikipedia.org/wiki/Directed_graph#CITEREFBang-JensenGutin2000) p. 19 in the 2007 edition; p. 20 in the 2nd edition (2009).

\*\*[Diestel, R., [Graph Theory, Electronic Edition](http://diestel-graph-theory.com/GrTh.html), 2005, p 12]\*\*

It is closely related to the theory of [network flow](https://en.wikipedia.org/wiki/Flow_network) problem: it represents an important measure of its resilience as a network.

* Net Robustness: **Robustness**, the ability to withstand failures and [perturbations](https://en.wiktionary.org/wiki/perturbation), is a critical attribute of many [complex systems](https://en.wikipedia.org/wiki/Complex_system) including [complex networks](https://en.wikipedia.org/wiki/Complex_network). n [economics](https://en.wikipedia.org/wiki/Economics), network robustness principles can help our understanding of the stability and risks of banking systems \*\*[Haldane, A. G.; May, R. M. (2011). "Systemic risk in banking ecosystems". *Nature*. **469**: 351–355. [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1038/nature09659](https://dx.doi.org/10.1038%2Fnature09659)]\*\* But in our specific case we can use the enginnering approach that uses it as a metric to evaluate the resilience of the infrastructure network that represent the stock market. \*\*[ Albert, R.; Albert, I.; Nakarado, G.L. (2004). "Structural Vulnerability of the North American Power Grid". *Phys. Rev. E*. **69**: 025103. [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1103/physreve.69.025103](https://dx.doi.org/10.1103%2Fphysreve.69.025103)]\*\*

The focus of robustness in complex networks is the response of the network to the [removal of nodes](https://en.wikipedia.org/wiki/Node_deletion) or links. The mathematical model of such a process can be thought of as an inverse percolation process. [Percolation theory](https://en.wikipedia.org/wiki/Percolation_theory) models the process of randomly placing pebbles on an n-dimensional lattice with probability p, and predicts the sudden formation of a single large cluster at a critical probability \*\*[ D. Stauffer and A. Aharony. Introduction to Percolation Theory. Tay-lor and Francis. London, 1994.]

https://en.wikipedia.org/wiki/Robustness\_of\_complex\_networks {\displaystyle p\_{c}}

<https://en.wikipedia.org/wiki/Percolation_theory>

<https://en.wikipedia.org/wiki/Connectivity_(graph_theory)>

<https://en.wikipedia.org/wiki/Menger%27s_theorem>

<https://en.wikipedia.org/wiki/Max-flow_min-cut_theorem>

* Diameter, net efficiency
* Distance + shortest paths: <https://en.wikipedia.org/wiki/Shortest_path_problem>

Concept used to calculate closeness centrality!

* Min spanning tree [https://en.wikipedia.org/wiki/Minimum\_spanning\_tree]

**Measures of centrality**

Centrality: answers the question: What characterizes an important vertex? These importance of a node can have different interpretations leading to different definitions of centrality. Given this two possible interpretations have been proposed: importance as related to a kind of flow across the network [Borgatti, Stephen P. (2005). "Centrality and Network Flow". *Social Networks*. Elsevier. **27**: 55–71. [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1016/j.socnet.2004.11.008](https://dx.doi.org/10.1016%2Fj.socnet.2004.11.008)] or as involved in the cohesiveness of the network.

indicators of **centrality** identify the most important [vertices](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) within a graph. Applications include identifying the most influential person(s) in a [social network](https://en.wikipedia.org/wiki/Social_network), key infrastructure nodes in the [Internet](https://en.wikipedia.org/wiki/Internet) or [urban networks](https://en.wikipedia.org/wiki/Urban_network), and [super-spreaders](https://en.wikipedia.org/wiki/Super-spreader) of disease. Centrality concepts were first developed in [social network analysis](https://en.wikipedia.org/wiki/Social_network_analysis), and many of the terms used to measure centrality reflect their [sociological](https://en.wikipedia.org/wiki/Sociology) origin [Newman, M.E.J. 2010. *Networks: An Introduction.* Oxford, UK: Oxford University Press.]. This will help to understand the topology/behaviour and dynamics of the studied system (central nodes provide the structure of the market maintaining it joined and if possible identify disease spreaders we can react.

Bonacich, Phillip (1987). "Power and Centrality: A Family of Measures". *American Journal of Sociology*. University of Chicago Press. **92**: 1170–1182. [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1086/228631](https://dx.doi.org/10.1086%2F228631)

 Benzi, Michele; Klymko, Christine (2013). "A matrix analysis of different centrality measures".[*arXiv*](https://en.wikipedia.org/wiki/ArXiv):[*1312.6722*](https://arxiv.org/abs/1312.6722)

Centralities are either Radial or Medial. Radial centralities count walks which start/end from the given vertex, while Medial…….

The centrality measures used in this work are:

https://en.wikipedia.org/wiki/Centrality

* + Degree centrality: Radial centrality measure. This one counts the number of walks of length one, so the number of first degree connections a node has. The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (flow studies) or cohesiveness measure.

In the case of a directed network (where ties have direction), we usually define two separate measures of degree centrality, namely [In Degree](https://en.wikipedia.org/wiki/Indegree) (count of the number of ties directed to the node) and [Out Degree](https://en.wikipedia.org/wiki/Outdegree) (ties that the node directs to others) these last are the ones computed TESLA for each node. Following a classic approach on Social Networks studies, when ties are associated to some positive aspects such as friendship or collaboration, In Degree is often interpreted as a form of popularity, and Out Degree as gregariousness, this means the links point to the nodes whose behaviour the source is following, source nodes are supposed to be affected by the nodes where the links point to.

* + - * Taking this into account In Degree of the nodes will be used as a measure to identify relevant stocks at which the rest are looking. Also we can use this directed links to study crash or recover spreads from hubs.
      * The definition of centrality on the node level can be extended to the whole graph, in which case we are speaking of *graph centralization [FORMULA]*
      * *Sector centrality????? (intra inter connections)*

Also, and given the way the interactions were recovered for each of the nodes in the same optimization problem enforcing sparsity we find, after some exploratory analysis, that the Degree out of all the nodes follow a similar distribution turning their average value as a good metric to measure the cohesiveness of a network and study its evolution over time with no need to look at the network.

[plots: degreeNodes, plot variation of out degree)

Intra-Inter Degree of Sectors?????

Eig. Centrality : Also radial centrality measure. **Eigenvector centrality** (also called eigencentrality) is a measure of the influence of a [node](https://en.wikipedia.org/wiki/Node_(networking)) in a [network](https://en.wikipedia.org/wiki/Network_(mathematics)). It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

https://www.math.washington.edu/~morrow/336\_11/papers/leo.pdf

* + Closeness centrality: In connected [graphs](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)) there is a natural distance metric between all pairs of nodes, defined by the length of their [shortest paths](https://en.wikipedia.org/wiki/Shortest_path_problem). The **farness** of a node *x* is defined as the sum of its distances from all other nodes, y, and its closeness was defined by Bavelas as the [reciprocal](https://en.wikipedia.org/wiki/Multiplicative_inverse) of the farness.

Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am*, **22**(6):725–730, 1950.

Thus, the more central a node is the lower its total distance from all other nodes. Note that taking distances *from* or *to* all other nodes is irrelevant in undirected graphs, whereas in directed graphs distances *to* a node are considered a more meaningful measure of centrality, as in general (e.g., in, the web) a node has little control over its incoming links (in this case a stock can’t control which other nodes are ‘looking’ at it??)

When a graph is not [strongly connected](https://en.wikipedia.org/wiki/Strongly_connected_component) (what is our case), a widespread idea is that of using the sum of reciprocal of distances, instead of the reciprocal of the sum of distances, with the convention {\displaystyle 1/\infty =0}1/inf=0:

Note that harmonic centrality is a most natural modification of Bavelas's definition of closeness following the general principle proposed by [Marchiori](https://en.wikipedia.org/wiki/Massimo_Marchiori) and [Latora](https://en.wikipedia.org/wiki/Vito_Latora) (2000)[[21]](https://en.wikipedia.org/wiki/Centrality#cite_note-marchiorilatora2000-21) that in graphs with infinite distances the harmonic mean behaves better than the arithmetic mean. Indeed, Bavelas's closeness can be described as the denormalized reciprocal of the [**arithmetic mean**](https://en.wikipedia.org/wiki/Arithmetic_mean) of distances, whereas harmonic centrality is the denormalized reciprocal of the [**harmonic mean**](https://en.wikipedia.org/wiki/Harmonic_mean) of distances.

* + Betweeness centrality: quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans (stocks) in a social network by[Linton Freeman](https://en.wikipedia.org/w/index.php?title=Linton_Freeman&action=edit&redlink=1)[[26]](https://en.wikipedia.org/wiki/Centrality#cite_note-freeman1977-26) In his conception, vertices that have a high probability to occur on a randomly chosen [shortest path](https://en.wikipedia.org/wiki/Shortest_path_problem) between two randomly chosen vertices have a high betweenness, it is computed:
    - For each pair of vertices (*s*,*t*), compute the [shortest paths](https://en.wikipedia.org/wiki/Shortest_path_problem) between them.
    - For each pair of vertices (*s*,*t*), determine the fraction of shortest paths that pass through the vertex in question (here, vertex *v*).
    - Sum this fraction over all pairs of vertices (*s*,*t*).

 Brandes, Ulrik (2001).[*"A faster algorithm for betweenness centrality"*](http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.2024)(PDF).Journal of Mathematical Sociology.**25**: 163–177.[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1080/0022250x.2001.9990249*](https://dx.doi.org/10.1080%2F0022250x.2001.9990249)*. RetrievedOctober 11, 2011*. Ref of the formula?

where {\displaystyle \sigma \_{st}} is total number of shortest paths from node s to node t and is the number of those paths that pass through v{\displaystyle v}. The betweenness may be normalised by dividing through the number of pairs of vertices not including *v*, which for [directed graphs](https://en.wikipedia.org/wiki/Digraph_(mathematics)) is (n-1)(n-2) (newton binomial??)

* + Percolation centrality: owever, a node may very well be centrally located in terms of betweenness centrality or another centrality measure, but may not be ‘centrally’ located in the context of a network in which there is percolation. Percolation of a ‘contagion’ occurs in complex networks in a number of scenarios. For example, viral or bacterial infection can spread over social networks of people, known as contact networks.
  + Disimilarity based o cent measures! Connected nodes doesn’t mean they are similar! (to do so we study which shared neighbours they have)

THIS one could be compared to the correlation?

* Node influence!
* Node dynamic importance: [Restrepo, Juan….PRL] In many situations it might be desirable to control dynamical processes that take place on networks. For example, in epidemic spreading, one would like to increase the threshold for epidemic transmission. In percolation, one might like to identify the key nodes holding the network together and protect them (e.g., in the transportation network orthe Internet) ordisrupt them (e.g.,inthe case of a terrorist network or pathogen protein network). Such strategies would greatly beneﬁt from a quantitative characterization of the effect of the removal of the different nodes or edges in the network. In many situations it might be desirable to control dynamical processes that take place on networks. For example, in epidemic spreading, one would like to increase the threshold for epidemic transmission. In percolation, one might like to identify the key nodes holding the network together and protect them (e.g., in the transportation network orthe Internet) ordisrupt them (e.g.,inthe case of a terrorist network or pathogen protein network). Such strategies would greatly beneﬁt from a quantitative characterization of the effect of the removal of the different nodes or edges in the network. use largest eigenvalue of adjacency matrix to calculate this, first check results of paper
* Crash and contagion study (gain/loss🡪 (Fixed, MDS, Sectors..)

Flow analysis: (GainLoss network) <https://en.wikipedia.org/wiki/Signal-flow_graph>: Signal-flow graph is based on the concept that the state of a node can be modelled by the previous state of the other vertices that point to it. Given this a node which In Degree is 0 is considered as an independent variable. We write a function for each node that processes the signals that it receives.

One variable I find interesting to add to the models is the previous state of the node itself, given that in this case we are studying dynamic networks it is consequent to consider the state of a node dependant on the previous state (in the signal-flow graphs loops are considered, not here, in these cases the same state of the node we are modelling is added to the function).

<https://en.wikipedia.org/wiki/Signal-flow_graph>

Dorf, Richard C.; Bishop,, Robert H. (2001). "Chap 2.-1: Introduction". [*Modern Control Systems*](https://www.site.uottawa.ca/~rhabash/ELG4152LN02.pdf) (PDF). Prentice Hall. p. 2. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [0-13-030660-6](https://en.wikipedia.org/wiki/Special:BookSources/0-13-030660-6).

**Analysis Sectors**

* (EVOL OVER TIME OF:) Centrality, Degree, Weights, inter and intra connectivity

**Analysis Stocks IMPORTANCE**

* Centrality, Degree, Weights

- Software used:

Matlab: Used to compute TESLA algorithm, which is built and ready to retrieve online. This script has been modified according to our needs, simplifying it, what makes it easier a further implementation of the whole ‘preprocessing + computation + analysis’ pipeline in R to generate a function that automatically recovers dynamic networks and analyses them (TESLA algorithm just finds the dependency strengths).

R: Used in the exploration/preprocessing stage and graph plotting and analysis thanks to the packages *igraph, network, sna* and *ggplot2*.

* **Experimental results**

- Data (characteristics, exploration and preprocessing). Different kinds of data and many cases in finance!! -> different approaches (changing parameters, analysing TESLA/correlations

As said before, TESLA estimates the dependency strength between nodes using time-series of the observed states of the nodes, hence the data used in this work correspond to time series of the stock values for the most important companies forming the British index: FTSE100.

First, and to get started data from 2009 and 2010 was taken from 85 assets that took part on the FTSE100 index on the mentioned years.

In order to study and analyse the impact of critical events data from January of 2015 until the end of July 2016 was gathered. In this case instead of focusing on the FTSE100 data from companies which Market Capitalisation is over 1B.

After some cleaning of the data to remove stocks which values were unreliable (missing values, many consecutive days no change) there remained 186 companies.

The FTSE100 consists of the largest 100 UK companies by full market value, given that the problem here is that the companies forming the FTSE100 change quarterly resulting in not homogenous data for large enough intervals, to solve this potential problem, data from all the stocks taking part on the FTSE100 in any time of the interval studied are retrieved (when data available during the whole interval). In a similar way, we noticed a couple of cases in which not all stocks had valuation probably corresponding to a vacation, this is solved by performing inner join (using only observations with values for all stocks).

Before moving on, it is important to notice that a large slice of these stocks are international companies, however, so the index's movements are a fairly weak indicator of the UK economy

A better indication of the UK economy is the FTSE250 since it contains bigger proportion of national companies. (study??) or FTSE350?

Gathering all these data hasn’t been trivial, since we will work with all stocks from FTSE market the best way to automatize this process is to connect to yahoo finance through R and iterate over the tickers of all the stocks in the British market. To do so first I stored all tickers from all stock markets and then filtered the dataset by country. It has been done this way in case there is enough time to study the behaviour of countries related to UK (Germany, France…) so we have already all tickers.

Since we assume that stocks with similar trend are likely to be influenced by the same underlying event, translating in the existence of a link between them, it is needed to calculate the daily returns. In addition to this, just using the stock values it is very unlikely to find links between companies with different order of values even if their trends are similar.

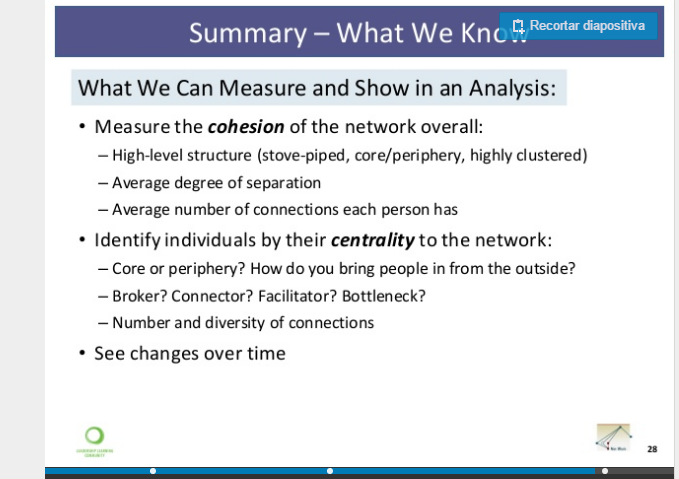
Once data is prepared we input it to TESLA.

First attempts it didn’t work at all, after simplifying some complex and useless features (for this problem) it started working.

After exploring and plotting first the results, I found out that the parameters for the indices i,i (node relation with itself) gave extremely high values (in this very case we are not interested in loops on the same node and it obscures the possible existing interactions with other stocks). Given the problem formulation on the TESLA paper [ref] we expect these values to be 0. In order to fulfil this condition some further modification of the algorithm was necessary.

- (updated)Algorithm + Analysis: Different analysis carried on after some exploration of the data and networks (few high degree nodes: highlight them, study them with analysis variables and last compare with news/historical data about those companies or general behaviour of market in those years (time series of stocks!)

The analysis performed to these networks focus on studying how the topology of the system changes on different situations, finding and studying the effects of the most important agents and understanding the relations between these stocks.



General topology

Definetely the recovered networks don’t follow a Scale Free distribution (baybe too small the first example used?) 🡪 small-world net!

- A **small-world network** is a type of [mathematical graph](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)) in which most nodes are not neighbors of one another, but most nodes can be reached from every other node by a small number of hops or steps;:

Specifically, a small-world network is defined to be a network where the typical distance *L* between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes *N* in the network, that is [CHECK FORMULA AND IF ITS POSSIBLE IMPLEMENT BY DELETING NODES!]

PROPERTIES AND REFERENCES!

RELATED TO ROBUSTNESS

https://en.wikipedia.org/wiki/Small-world\_network

The first networks computed showed a star distribution in which there is one or few central nodes (hubs) that connects to (almost) all rest of the ‘weak’ nodes, usually only connected to the hubs, setting the background net on which the whole system is built. An advantage of the star topology is the simplicity of adding additional nodes as well as its reliability since because if one node or its connection breaks it doesn’t affect the other nodes and their connections. On the other hand the hub represents a single point of failure, the less hubs the system have the less robust the system is.

First thing we can notice after having a look to the financial networks is that it has similar characteristics than a scale-free distribution but doesn’t follow a power law, instead there is usually a small number of high-degree nodes (hubs) while the rest follow quasi uniform values.

Robustness???

Pathways??

Weak or strong nets when facing contagion?? (notes)

GainLoss matrix: contagion??? :::::: use different shape for financials!!!!!! (try to study if they represent riskier nodes):::::::: study this mathematically (so find the proportion of adjacent nodes that gets the same state [+/-] that the source node had the previous epoch taking into account the sign of the interaction (so if the interaction is neg: the nodes would need to have different sign), and measure the average abs(weight) of the nodes that joined them (esto es porque las interacciones negativas ya las hemos tenido en cuenta) so we can check the viability of being the contagion source)-> maybe more useful in high freq analysis)

Also: sum incoming weights of node i (taking into account the state (sign) of the sources): sum(weightsIn(j)\*GainLoss(j)) (normalize??) j represent every node but the one we are studying. After this we can compare with the Gain/Loss value of the node i (in the same period? The next period??) and try to find a correlation: or a model. If we can model and predict accurately enough (error, boxplots…) the gain or loss of a stock given the incoming weights of its connected nodes this could be extremely helpful for companies to prevent situations and react before to minimise the loss or maximise the gains. \*\*\*try different approaches: maybe with the averaged weight, or taking into account other factors as the previous Gain/Loss value of the node I, or its value on the stock market, or weighting the weights by the companies’ value (CP)??(4M, 2.1B, 856M…) of the source nodes: sum(weightsIn(j)\*GainLoss(j)\*CP(j))/sum(CP(j))

And maybe multiplying or using CP(i).

CENTRALITY MEASURES

Information about the relative importance of nodes and edges in a graph can be obtained through [centrality](https://en.wikipedia.org/wiki/Centrality) measures

"Importance" can be conceived in relation to a type of flow or transfer across the network. This allows centralities to be classified by the type of flow they consider important[ Borgatti, Stephen P. (2005). "Centrality and Network Flow".Social Networks. Elsevier.**27**: 55–71.[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1016/j.socnet.2004.11.008*](https://dx.doi.org/10.1016%2Fj.socnet.2004.11.008).] "Importance" can alternately be conceived as involvement in the cohesiveness of the network.

We can try to model the general cohesion of a net with the historic market volatility (of previous state). By considering the avg degree (weighted?)

Degrees: total, in out????

Try to find relationship between adjacent networks’ hubs (predict next hubs given the previous ones???)

(or predict prob of turning hub given its previous weights in(norm?), sector, values, gain/loss: Decision trees?: Input=previous net info dataframe(nodes values, edges values, sectors, gain/loss, company values, avg(stockValue)…). Output=hub? Gain/loss?-> mejor, es un valor directo del stock market, los hubs son calculado por TESLA

One important characteristic of the recovered networks is that they are directed given that the adjacency matrices computed by the TESLA algorithm returns the relation of each one of the nodes (independently) with the rest on a single inference procedure. This results on not (necessary) reciprocal interactions. In this case it is necessary then to study out (connections of the computed node with the rest) and in (connections of the other computed nodes relating this one) degrees of the nodes.

Degree out: Number of computed connections that emerge from the studied node when applying TESLA to find its relation with the rest.

Having a look to Out Degree tables we can observe that they are quite homogenous in general (excluding some outliers we will study further as possible hubs), this is given by TESLA since the time-series used for each iteration (stocks) are almost the same (every node but the one we are studying) and they share constraints. Although the Out Degree is usually similar in all nodes in a given network, when comparing these values in different time stamps we find that they vary [buscar context cuando desconectada], pointing the potential of this variable to measure market’s global cohesion.

[[First of all we would compare this value with the degree in, is it a hub.in?, then the time series of the stock….]]

Degree in: Number of computed connections that reach a given stock.

On these histograms we can see a completely opposed distribution of the Out Degrees. Since this degree gives us the number of times a given stock appears as related to the rest when applying TESLA to them we can just use it as a relevance/centrality measure. The higher the degree, the most relevant is the stock for the market so it will have more effect on it (leading the system, spreading a contagion, connecting the individuals and setting a background net over which the market structures…) finish:::: hubs:::: dominating vertex (n-1 connections)

(degree variation over time: sectors and general) In the example case we can observe how the degree of all sectors decrease at the same time indicating a reaction to ?????

Also here we can check how although the financial sector has bigger degree, the stocks belonging to the utilities sector represent stronger hubs as they have more average number of edges attached. Probably the Financial sector is more important as a whole (maybe highly intraconnected approaching the behaviour a supernode itself?) but the average stock in the Utilities sector is stronger than the ones in Financials. This could be interesting: degree of a sector vs avg degree of stocks from sector

Eigenvector centrality

Betweenness centrality

Katz centrality???

Weights

Values and distributions by individuals and sectors

Intra and inter sector connectivity

In order to study importance of different sectors over time we can compute sector degrees and weights.

Also to know how much a sector is isolated or interacts with the market we should study its intra and interconnectivity with other sectors by measuring the proportion of connections a given sector have between its stocks and with different sectors.

The bigger the intraconnectivity compared to the interconnectivity the more isolated a given sector will be while the bigger the interconnectivity the more active it is.

Layouts (physic models, MDS, random fixed…)

Different information from different approaches.

Main ones: kk layout (particles and springs) to study how the general system evolves and random/circular fixed to focus on how the links changes over time and study flows or contagions.

MDS gives us the similarities between the stocks (using Euclidean distance and computed for each of the epochs to compare with the network)

Explain layouts

TODO: Network grouped by sector. Kk-means. Network coordinates = pca???

Show plots and explain encodings: size = degOut+degIn/10 (so we can see the hubs double than the rest but still can see them), vertex colors=sectors, edge colors: edge in, clockwise flow….

Basically enumerate all analysis carried on and why (based on network analysis, on other papers or made up after some exploration) and how they relate to each other, which results we can extract from all data analysed together…)

* **Conclusions**
  + Comparison with other works
  + Further work (next month, next 10 years)
  + Conclusions

Modelling the stock market as a network and using their relationships to predict future behaviours as gain or loss of the stock values, future values themselves?, has resulted ………………(comment errors), also comparing with the results given using the correlation matrix values instead of the TESLA interaction matrix values (same tasks, measure of error from both of them and compare!!)

* Reachability??: connectivity
* -https://en.wikipedia.org/wiki/Strength\_of\_a\_graph???

