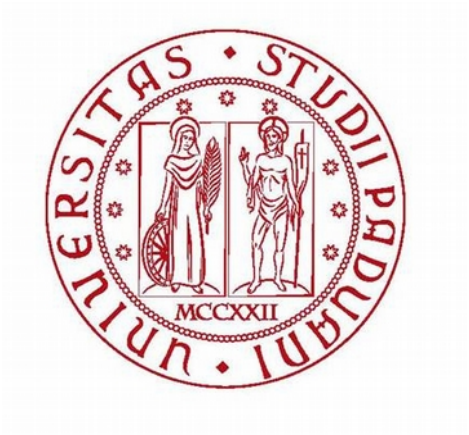

UNIVERISTÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

TESI DI LAUREA MAGISTRALE

Sleeping beauties and the citation dynamics in the network of scientific papers

Supervisors:
Filippo Simini
Naoki Masuda

Internal Supervisors:
Amos Maritan
Samir Suweis

Candidate:
Fabio Peruzzo

Academic Year 2014-2015

Contents

1 Network science: historical remark

Network science is nowadays an important part of the big family of complex systems, and finds the main reason of its popularity in its versatility: from biology to economy, from social to computer sciences, it can be applied to a very heterogeneous number of problems. The main ingredients are very simple, just points linked together by lines or arrows.
Even though it is now so widely used, network science does not have a very long history: its development as an independent branch of probabilistic analysis is relatively recent.
What was probably the first problem formulated in terms of nodes and edges was the so called 'Seven Bridges of Konigsberg', by Euler in 1736: the Prussian city of Konigsberg, now in Russia, is built around two islands that emerge from the waters of a river called Pregel. These islands are linked together and to the rest of the city by seven bridges: the problem proposed by Euler was to devise a walk around the city that would cross each bridge once and only once. The difficulty was to prove if such a path existed with mathematical rigor (Fig 1).
Since the path within an island or the land is totally irrelevant, Euler simpli-

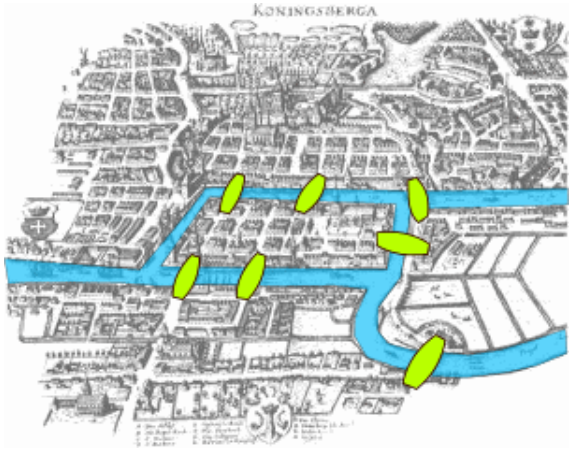

Figure 1: Image of the city of Konigsberg at the time of Euler

fied the analysis considering only dots for each of the places he had to reach (each land connected by the bridges), and indicated the connections between them only with lines. Next, Euler observed that (except at the endpoints of the walk), whenever one enters a vertex, one has also to leave it using a bridge. In other words, during any walk in the graph, the number of times one enters a non-terminal vertex equals the number of times one leaves it.
In modern terms, Euler demonstrated that the existence of a solution de-

pends on the degrees of the vertices, i.e. on the number of lines touching the point.

After this first and rudimentary usage of nodes and edges, not much was done in order to develop these ideas for almost two centuries.
In 1930 we find another pioneristic application of graphs, this time to represent the social structure of a group of elementary school students (fig 2).

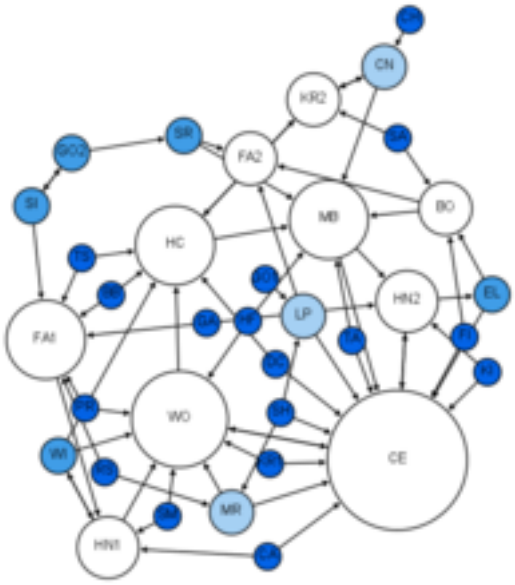

Figure 2: Sociogram of a group of student as represented by Jacob Moreno (Ref [1])

Some interesting features of this network were, for example, that boys tended to be friends with boys and girls with girls, with only some rare exceptions, but, again, this analysis was not done with rigor or introducing any specific mathematical tool.
It is only with the works of the Hungarian mathematicians Paul Erdos and Alfred Renyi that graph theory starts to be properly analyzed as a branch of probabilistic theory, introducing the concept of random graphs (Ref [2]).
Consider a set of n isolated vertices. A first random model could be the one in which we start adding successive edges between these nodes at random until a certain total number of edges m is achieved.
In another very important random model, initially proposed by Edgar Gilbert and denoted $G(n, p)$, every possible edge occurs independently with probability $0 < p < 1$. In this framework, obviously, the probability of obtaining

any specific configuration is:

$$p^m \ (1-p)^{(N-m)}$$

where

$$N = \binom{n}{2}$$

Modifying this prescriptions, Erdos-Renyi proposed a new model (called $G(n,m)$), in which all the networks that have the same number of edges $(m)$are equiprobable.
More precisely, $\binom{N}{m}$ are all the possible configurations of the network, and each of these occurs with probability

$$1/\binom{N}{m}$$

However, what is probably the best known random network is the *exponential random graph* (Ref [3]).
Its basic assumption is that the observed structure of the network can be explained using only some nodal attributes, together with some parameters and a normalization function:

$$P(y) = \frac{e^{\theta s(y)}}{N(\theta)} \tag{1}$$

where $s(y)$ can be any function of the nodes of the network (e.g. the sum of the number of links that arrive at each of them), N is a normalization function and θ a parameter.
It is immediate to note that this probability distributions resemble the canonical distribution in statistical mechanics.
A very interesting and useful tool that was introduced in those years is the *adjacency matrix* Φ: every index represents a node, and each single entrance $\Phi_{ij} = 1$ if there is a link from node i to node j, otherwise $\Phi_{ij} = 0$.
Almost all the interesting cases in which network science is used are systems that are changing with time, and in which we are trying to predict some future characteristics: recently a lot of effort has been put on trying to understand evolving networks. Not without surprise, it has been found that many apparently unrelated systems manifest a lot of similarities.

2 Definition and coefficients in network analysis

Generally speaking, a network is a collection of points (nodes) that are linked together by lines(edges). These edges can be of different forms and types: in particular, we say that a network is

- *direct* if its edges have a defined direction, i.e. the relation involving the extremes goes from one node to the other. Normally, the edges are represented as arrows
- *indirect* if in the relation among nodes there is no specified direction (edges are simple lines between them)

Obviously, using one or another depends on the system and the application we are dealing with.
In order to be more quantitative we also need to introduce coefficients:

- *Density D*: ratio of the number of edges E to the number of possible edges of the network:
$$D = \frac{2\ E}{N(N-1)}$$
- *degree of a node k_i*: If i is a node in an indirect network, its degree is the number of arriving edges.
If the network is direct, we will separate the in degree k_i^{in}, i.e. the number of edges that arrive in i, from the out degree k_i^{out}.
- *Average path length*: calculated by finding the shortest path between all pairs of nodes, adding them up, and then dividing by the total number of pairs
- *Diameter of the network*: the longest of all the calculated shortest paths in a network, i.e. the longest distance between two pair of nodes in the network, calculated in number of nodes one has to cross.
- *Clustering coefficient C_i*: the clustering coefficient measures how many of the neighbors of a node are linked together. More precisely, the clustering coefficient of node i is calculated as:
$$C_i = \frac{2e_i}{k_i(k_i-1)}$$
where k_i is the degree of the node and e_i the total number of edges in the sub graph made keeping only the neighbors of i (its maximum value is, of course $\binom{k}{2}$).

- Connectedness: there are various ways to analyze how connected a network is, and all deal with the average number of paths between nodes. For example, a very common problem on network is to find when the giant component arises, i.e. when the system reaches the percolative threshold and almost all the nodes are connected to each other by links.

- *node centrality*: Since 'centrality' is a concept that depends on the type of network, obviously there is no general definition of this index. However, some quantities are more used than others. For example, *Betweenness centrality* is a measure of a node's topological centrality inside the network. If i is the node, it is defined as:

$$g_i = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

  where σ_{jk} is the total number of paths between nodes j and k, and $\sigma_{jk}(i)$ is the number of these that pass through i.
  Another coefficient is *eigenvalue centrality*, that measures the importance of a node in the network: assigning relative scores to all the nodes, connections to high-scoring nodes contribute more to the score of a node than low-scoring nodes. So, if i is the node and Φ is the adjacency matrix, we have that the centrality coefficient for i, x_i, is calculated as:

$$\sum_{j \in G} \Phi_{ij} x_j = \lambda x_j$$

  where $j \in G$ means that we are considering all the nodes of the graph, and λ is a constant. There may be a lot of values for λ that admit eigenvectors for the matrix Φ, but only one of these has all the x_i non negative numbers (it is a consequence of Perron-Frobenius theorem).

There are many other coefficients with which one can classify nodes and networks, but in the rest of the thesis we will only use very few of them, so I'm not going to include any other definition.
In the following section I'm giving some examples of real systems that can be represented as networks, and I'm going to mention some of the most important results achieved through this analysis.

2.1 Examples of real networks

- **The world wide web**: the nodes of this network are web pages, i.e. data that may contain images, video and words. The edges are hyperlinks that point from one page to another (the www is a direct network).
  In 1999 it was discovered that both the distribution of in degree and out degree of edges follow a power law over several orders of magnitude (Albert, Barabasi, [5]):
  $$P(k) \sim k^{-\gamma}$$
  with $\gamma_{out} \sim 2.4$ and $\gamma_{in} \sim 2.1$. Moreover, using finite size scaling, it was also demonstrated that it displays the 'small world' property: the average distance between two nodes is ~ 19, a surprising property if one considers that the total number of pages is of order 500 millions!

- **Internet**: Internet is a network of physical links between computers and other telecommunication devices. Its topology can be studied at two different levels: in the first level, each node is a router, and edges are physical connections between them. In the second, hundreds of computers and routers are represented together, by a single node, and a link exists if there is at least one route that connects them.
  Studies conducted on it have revealed that the degree distribution follows a power law in both cases.

- **Science collaboration graph**: the nodes are scientists and two are connected if they appear as co author in at least one article.
  In 2001 Mark Newman in the work in ref [6] analyzed papers in databases of physics, bio medicine, high-energy physics and computer science published in a five-year window, from 1995 to 1999.
  What he found was a small average path length but a high clustering coefficient. Moreover, the degree distributions were not simply power law, but experienced an exponential cutoff at a certain point:
  $$P(k) \sim k^{-\tau} e^{-k/z_c}$$
  where τ and z_c are constant that depend on the specific field of research, but the functional form is the same for all the subjects that were studied (fig 3).

- **Cellular networks**: The nodes of this network are the substrates (such as ATP, ADP, H_2O) and the edges represent the predominantly directed chemical reactions in which these substrates can participate.

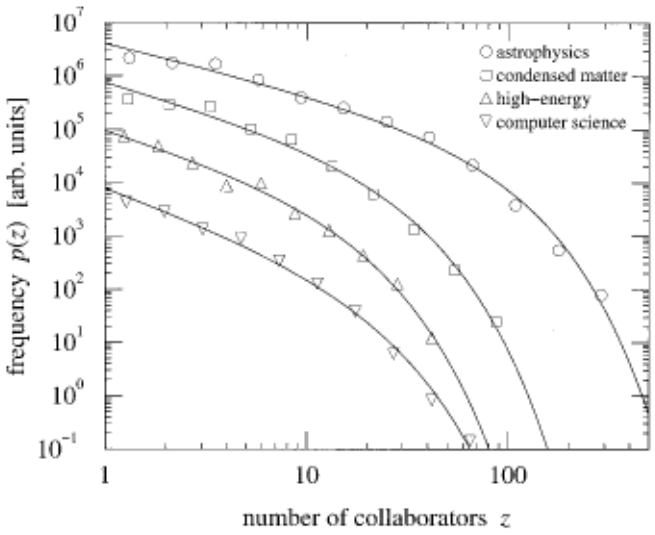

Figure 3: Distributions of the number of collaborations, divided for subject. each of these can be described using a power law distribution with an exponential cutoff.

The average path length was found to be approximately the same in all organisms, with a value of 3.3, and again the degree distribution follows a power law.
Another similar network describes protein-protein interactions, where the nodes are proteins and they are connected if it has been experimentally demonstrated that they bind together. Again, the degree distribution follows a power law with an exponential cutoff:

$$P(k) \sim (k + k_0)^{-\gamma} e^{-(k+k_0)/k_c}$$

- **Ecological networks**: Food webs are used regularly by ecologists to quantify the interactions between species: in this type of systems, nodes are species and the edges represent predator-prey relationships between them. In the analysis one discovers that even though species may differ from one ecosystem to another, they all are three or fewer edges from each other, and, again, degree distribution is consistent with a power law.

3 Citation network: definition and models

The subject of the following analysis will be the network of scientific citations. Every time an author starts a research work, he searches for information on what has already been discovered on that specific topic, and from those ideas he adds his ones in order to find something new.
When it comes the time of writing the paper the names of this previous works, that have been useful in the new research, are put in a *reference list* that somehow takes into account the credit every paper owes to previous papers. Moreover these lists are useful also to put the reader in the condition of understanding what is being expressed in the paper: single concepts that may be hard to understand can be found treated in more detail in previous works.
So reference lists' aim is to provide both credit to colleagues and to put a paper in a specific research scenario, giving the reader information on how to deepen into it in case he is interested.
In the network of scientific citations every paper is represented by a node, and one draws a direct edge between A and B if A cites B, i.e. if B is in the reference list of A.
One obvious property of this system is that it is mostly acyclic, since papers are published with a precise time ordering.
During the last 50 years a lot of effort has been put to better understand this system. It represents a unique opportunity to analyze the spreading of information and innovation between people.
The first thing one notices analyzing publications is that the yearly amount of papers grows exponentially (Fig 4), together with the average number of papers in reference lists (Fig 5) .

A large number of models of citation dynamics have been built, and in the next section I'm going to analyze some of the most significant.

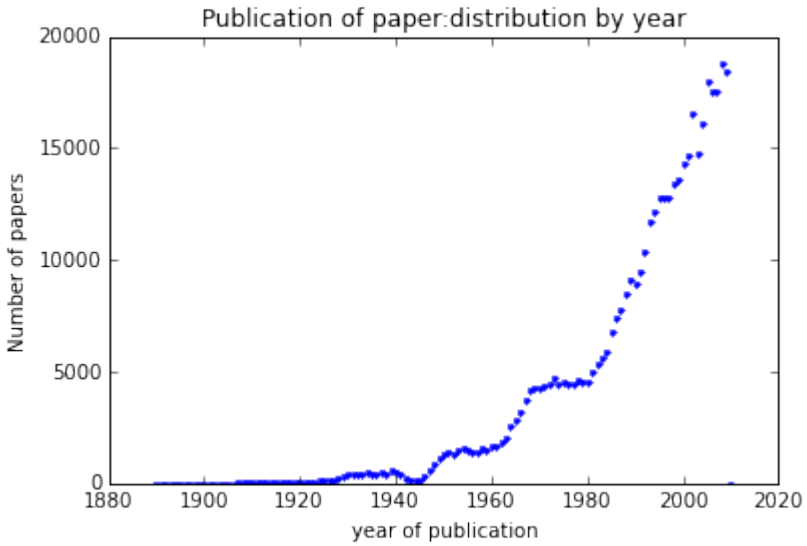

Figure 4: Distributions of publications per year: apart in the period of the war, the number of publication has never decreased, and, apart in two specific periods in which APS had decided to limit its number of publications, when it is left free this number grows exponentially

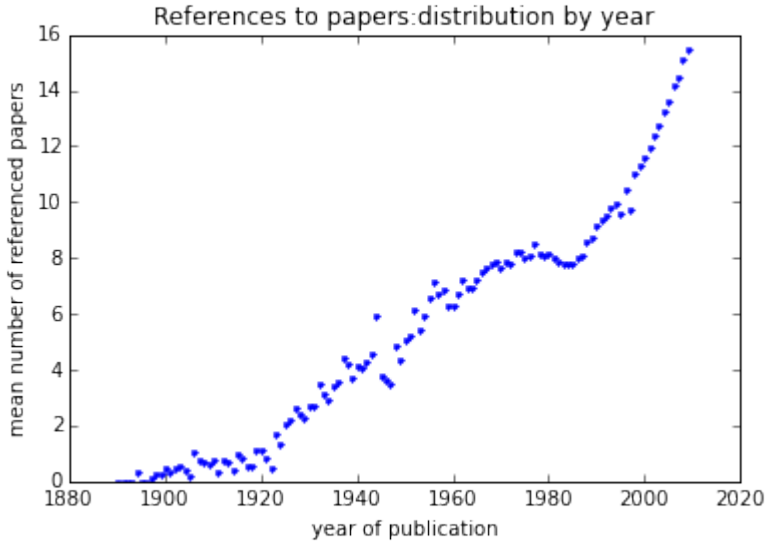

Figure 5: Distributions of the mean number of references per year. The growth is clear (the only exception was in 1980, due to a precise publication choice by APS), and normally is described via an exponential

3.1 Cumulative advantage

In 1965 professor Derek de Solla Price published what is probably the first quantitative study on citation dynamic [7].
In his work he noted some striking features: the distribution of papers' citations appeared to be very skewed on small numbers, but on the other hand it had also a very 'long tail' consisting of few very popular works. If P_k is the fraction of papers cited exactly k times, this value decreased with the increasing of k as $P_k \sim k^{-\alpha}$, with α a constant (a distribution known as Pareto Tail or power law).
In another paper, Price proposed a mechanism on how this type of distribution can arise, called *cumulative advantage*: papers that have gained more citations in the past are more likely to get new ones in the future.
In 1999 Barabasi and Albert [5], studying the distribution of links between pages in the *world wide web*, independently proposed a similar process of link gaining by web pages, now calling it *preferential attachment*.
We start considering that each paper that is published on average cites m other articles (i.e. the reference list has m papers), chosen in proportion to the number of citation k they already have plus a positive constant r (necessary to ensure that new publications can still receive citations). So we take all the indegrees of papers, sum them up and consider as a probability for each paper the ratio between its actual in degree and this total number of citations.
This mathematical problem can be solved exactly in the limit of large number of papers using a master-equation method introduced by Simon [4].
What is found is that, if p_k is the fraction of papers with exactly k citations,

$$p_k = \frac{B(k+r, \alpha)}{B(r, \alpha - 1)} \tag{2}$$

where

$$B(a, b) = \frac{\Gamma(a)\ \Gamma(b)}{\Gamma(a+b)},$$

Γ is the standard gamma function

$$\Gamma(n+1) = n!$$

and

$$\alpha = 2 + \frac{r}{m}. \tag{3}$$

Since for large values of its first argument $B(a, b) \sim a^{-b}$, the tail of the distribution of k (since we are considering $r \ll k$ in this limit) is

$$p_k \sim k^{-\alpha}$$

3.2 The first mover advantage

This is how *preferential attachment* is explained by means of Price's argument, but using the same ideas we can go a little bit further, and calculate the full distribution of citations as a function of its date of publication.
As pointed out by M. E. J. Newman in his work 'The first-mover advantage in scientific publication' [8], preferential attachment also implies a variety of other features, such as a strong *first mover advantage*: the first papers published in a certain topic should experience far higher rates of citations than those that come after them.
To prove this, he started from defining a 'time' variable t such that the i^{th} paper published has $t = i/n$ (t has no relationship with actual time, but gives a specific order of appearance of papers), and n is the total number of papers.
We define the density function $\pi_k(t, n)$ such that $\pi_k(t, n)dt$ is the fraction of papers that have been cited k times and that were published in the interval from t to $t + dt$.
In the limit of a large number of papers, if we define $\pi_k(t) = \pi_k(t, \infty)$, after some tedious calculations we get:

$$(k + r)\ \pi_k(t) - (\alpha - 1)\ \frac{d\pi_k}{dt}\ t = (k - 1 + r)\ \pi_{k-1}(t)$$

with the convention $\pi_{-1}(t) = 0$ and $\pi_k(0) = \delta_{k,0}$, and α defined in eq 3.
The solution of this equation is:

$$\pi_k(t) = \frac{\Gamma(k + r)}{\Gamma(k + 1)\Gamma(r)}\ t^{r/(\alpha-1)}(1 - t^{1/(\alpha-1)})^k; \tag{4}$$

and from this we can calculate the average number of citations $\gamma(t)$ a paper receives as a function of its time of publication:

$$\gamma(t) = \Sigma_{k=0}^{\infty} k\ \pi_k(t) = r\ (t^{-1/(\alpha-1)} - 1) \tag{5}$$

We can notice how for $t \to 0$ this value becomes arbitrarily large, meaning that early published papers are expected to receive more citations than the ones published later on (note that α is always bigger than 2 by definition). This is what is called *First mover advantage*

To test this hypothesis, Newman considers papers within a single specific research field, searching for data that describe it from its earliest foundation. This can be a very hard task for the majority of topics, so Newman limits his analysis to few specific subjects, such as network theory.

Starting from the first five early (and very well cited) papers in this field, he builds the network with all the papers that cite them (excluding review articles), adding the ones that cite these ones and so forth. The resulting data set contains 2407 papers spanning a ten years period from June 1998 to June 2008.

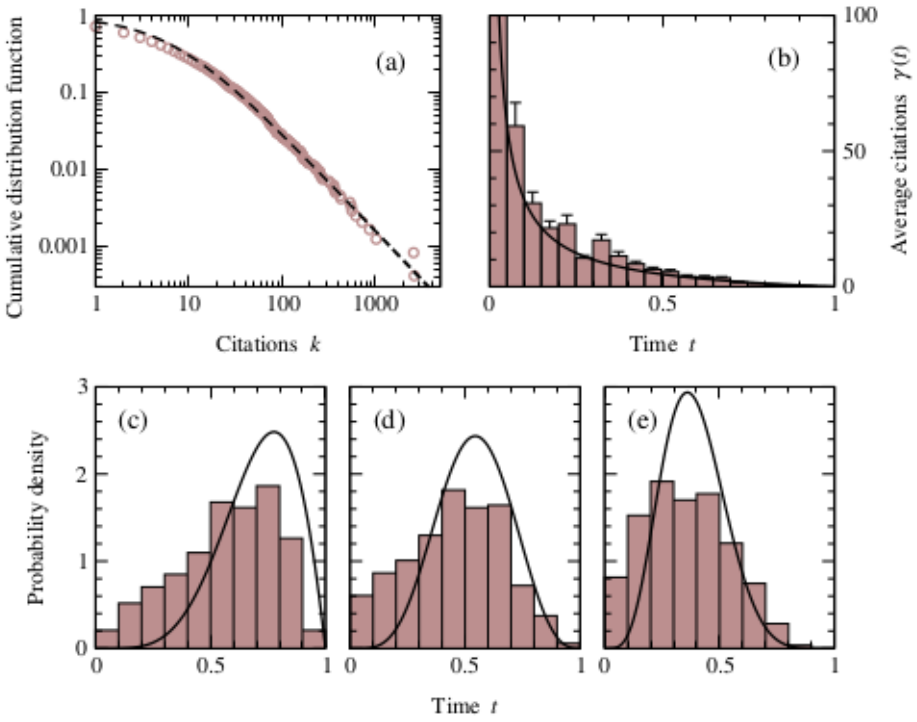

Figure 6: Figure a compares real cumulative distribution of papers on network science with the predicted one. Figure b shows the mean number of citations received by papers as a function of time from beginning (t = 0) to end (t = 1) of the covered period. Figures (c), (d) and (e) show the probability that a paper with a given number of citations is published at time t for papers with (c) 1 or 2 citations, (d) 3 to 5 citations, and (e) 6 to 10 citations.

.

Figure 6 a shows the distribution of citations of these papers, and as you can see the exponential fit is remarkably good. From it, we can extract values of r and a as seen in Eq 2.
Figure 6 b shows the average number of citations received by papers, where time is intended to be, as already mentioned, in terms of publication order. The solid line shows these values as predicted by eq 5: the agreement is quite

good, and shows a strong first mover advantage.

However, *preferential attachment* alone is far from describing all the features of citation dynamics. Figures 6 c-e report the distribution of citations at different times. Theoretical and actual values are quite different: there are much more papers published at early times in each degree range and fewer around the peak, meaning that not all the papers in the early period are benefiting from the first mover advantage.
Moreover, from a more accurate analysis of the data set one can notice the appearance of well cited papers also relatively late. In order to quantify this phenomenon, suppose we are interested in papers published after a certain t_0. Their distribution in the *preferential attachment model*, denoted $p_k(t_0)$ can be calculated simply integrating 4:

$$p_k(t_0) = \frac{1}{1-t_0} \int_{t_0}^{1} \pi_k(t) dt$$

This way, what we find is that $p_k(t_0)$ follows a truncated exponential behavior, but comparing these results to real data we find poor quantitative agreement (Fig 7).

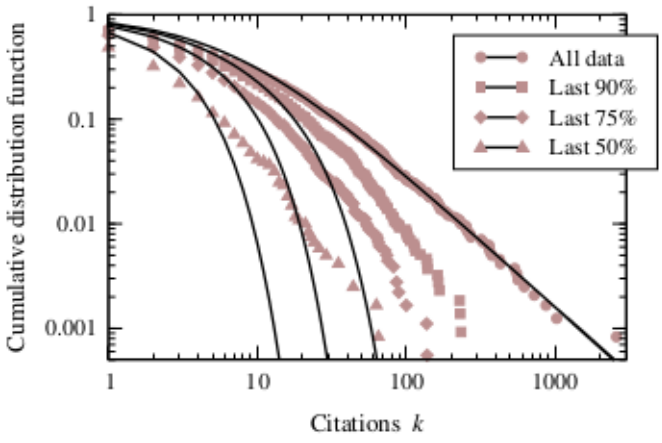

Figure 7: cumulative distribution function for subsets of papers in the data set. The top curve represents the whole data set, while the others represent the most recent 90%, 75% and 50%

Many other problems also arise as soon as we start considering other research topics. For example, if we take all the papers that deal with strange matter, we can see that the exponential behavior is lost even considering all the data (Fig8).

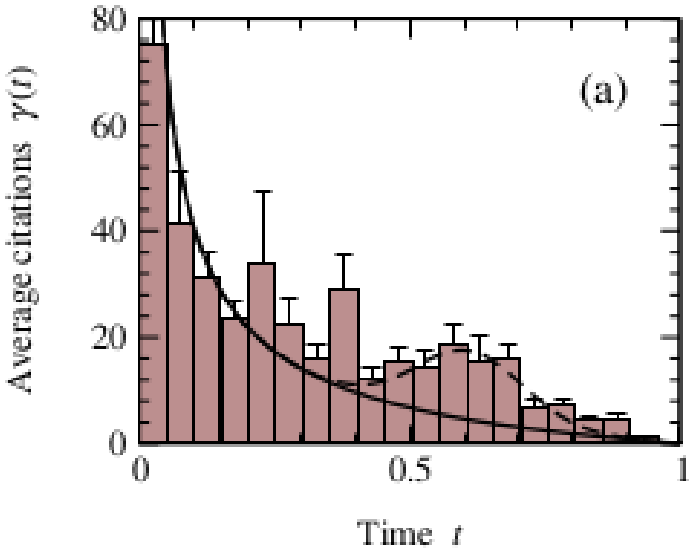

Figure 8: total distribution of citations for papers on Strange matter. Around $t \sim 0.6$ we find a new smaller peak, and the exponential fit does not describe the behavior of data

.

This is probably due to the fact that the interdisciplinary nature of physics makes it very difficult to understand where a subject ends and where another starts.

3.3 Random graphs for directed acyclic networks

In 2009 a new model was proposed by Karrer and Newman (Ref [9]) who provides a basic theory for directed acyclic networks, i.e. exactly the type of systems that best represents citation dynamics.
We can start by considering graphs with a fixed nodes degree sequence: if N are the vertices, with $i = 1, ..., N$ an index on them, let be k_i^{in} and k_i^{out} their fixed degree sequence.
Since the network is time ordered, edges are allowed to run only from vertices with higher to vertices with lower i value, and this constraint enforces the acyclic nature of the network: we can have an edge running to vertex i from vertex j only if $i < j$.
This way not all the realizations of the network are possible since there are precise conditions one has to satisfy. For example the sum of the in-degrees of all vertices must be equal to the sum of the out-degrees, because every edge that starts somewhere ends somewhere.
If m is the total number of edges in the network, we must have that:

$$\sum_i k_i^{in} = \sum_i k_i^{out} = m$$

Other conditions are, for example, that the first node ($i = 1$) must not have edges departing from it, (there is no node to which it can point).
We can consider the number of edges that pass over a specific node i and the ones that do not match any other stub before i:

$$\mu_i = \sum_{j=1}^{i-1} k_j^{in} - \sum_{j=1}^{i-1} k_j^{out}$$

i.e. the number of in going stubs below vertex i that are available to attach to outgoing stubs at i and above. One obvious requirement is that

$$k_i^{out} < \mu_i$$

For convenience, we will define also:

$$\lambda_i = \sum_{j=1}^{i-1} k_j^{in} - \sum_{j=1}^{i} k_j^{out}$$

and our condition that $k_1^{out} = 0$ can be written as $\lambda_1 = 0$, and $k_N^{in} = 0$ as $\lambda_N = 0$ (there is no node after N).
It can be demonstrated that these two conditions, together with $\lambda_i > 0$,

are necessary and sufficient for our degree sequence to properly represent a direct acyclic graph. The quantities μ_i and λ_i have a simple geometric interpretations: if we make a cut in our graph between vertices i and $i-1$, the quantity μ_i is the number of edges that cross the cut, or the number flowing from higher to lower vertices. For this reason, we call μ_i the flux at vertex i.
The quantity λ_i is equal to the number of edges that flow 'around' vertex i, meaning the number that run from vertices above i to vertices below. We call this quantity the excess flux at vertex i.
Considered all these quantities, Karrer-Newman's model takes the in stubs and out stubs at each vertex and tries to match it with all the other stubs of the network. The ensemble of all such matchings appearing with equal probability, constitutes the model.
There are some subtleties to this operation: matchings of stubs are not in one-to-one correspondence with network topologies. If we take a matching and simply permute the labels of the out-stubs at a single vertex i, we produce a new matching corresponding to the same topology.
The number of distinct permutations to arriving edges is $k_i^{out}!$, and the total number of permutations will consequently be:

$$\prod_i k_i^{in}!\ k_i^{out}!$$

there is another complication: if multiedges are present in the graph, then some configurations are over counted, and in order to avoid this mistake we must reduce our number by a factor:

$$\prod_{i<j} A_{ij}!$$

where A_{ij} are the elements of the generalized adjacency matrix(instead of only being 0 or 1, the elements of the matrix indicate the number of edges that run from vertex i to j).
One of the most fundamental properties of our model is the expected number of directed edges between any two vertices i and j. If P_{ij} is such a quantity, we find that:

$$P_{ij} = \frac{k_i^{in} k_j^{out}}{m} f_{ij}$$

where

$$f_{ij} = m\ \frac{\Pi_{l=i+1}^{j-1} \lambda_l}{\Pi_{l=i+1}^{j} \mu_l}$$

Another very interesting quantity is assortativity, i.e. vertex correlation on some variable.
Consider a quantity x defined on all vertices i. The network is said to be assortative with respect to x if edges tend to connect vertices with similar values of x, high with high and low with low, and can be calculated with a standard Pearson correlation coefficient r.
In a direct network we can consider more complicated forms of correlation, even involving two quantities x and y: a possible definition by means of the adjacency matrix could be

$$r = \frac{1}{\sigma_X \sigma_Y} \left[\frac{1}{m} \sum_{ij} A_{ij} x_i y_j - \mu_{in} \mu_{out} \right]$$

where

$$\mu_i^{in} = \frac{1}{m} \sum_i k_i^{in} x_i \qquad \mu_i^{out} = \frac{1}{m} \sum_i k_i^{out} x_i$$

and

$$\sigma_X^2 = \frac{1}{m} \sum_i k_i^{in} x_i^2 - \mu_{in}^2 \qquad \sigma_Y^2 = \frac{1}{m} \sum_i k_i^{out} y_i^2 - \mu_{out}^2$$

Conventional random models show no assortativity, but in the random acyclic case we can have non zero assortativity with respect to some quantity x, e.g. vertex degree $x_i = k_i^{in}$ and $y_j = k_j^{out}$.

A slight modification of what has just been defined is the *random directed acyclic graph with independent edge probabilities*, in which, rather than fixing the degree of each vertex, we fix only their expected values.
Starting with an empty graph of N vertices we generate for each pair of vertices i and j, with $i < j$, a Poisson distributed number with mean P_{ij} and place that number of edges between i and j, pointing from j to i. The resulting network trivially has the same expected number of edges between every vertex pair as the network generated by our first model with the same degree sequence, but the edges are now, by construction, independent.

3.3.1 Wu Holme model

In 2009 another paper was published (Ref [10]) by Wu and Holme that, trying to use these models, highlighted some problem in dealing with triangles, i.e. edges that connect three papers so that $a \to b$, $b \to c$ and $a \to c$.
In fig 9 I have reported the results of Wu and Holme analysis of triangles in Karrer Newman's model (KN model in the following).

Together with the simple KN model, they also used an extended version of it in which 'when a new vertex enters the network, rather than randomly matching all its out-degrees with those in-degrees among the existing vertices, after first matching one out-degree randomly with an in-degree belonging to an older vertex w (like the KN model), we let as many of the remaining arcs as possible to come from neighbors of w (and after that, also the neighbors of its new neighbor).
Note that, by definition of the KN model, network size N and degree sequences (in and out degrees) are identical to empirical data.'

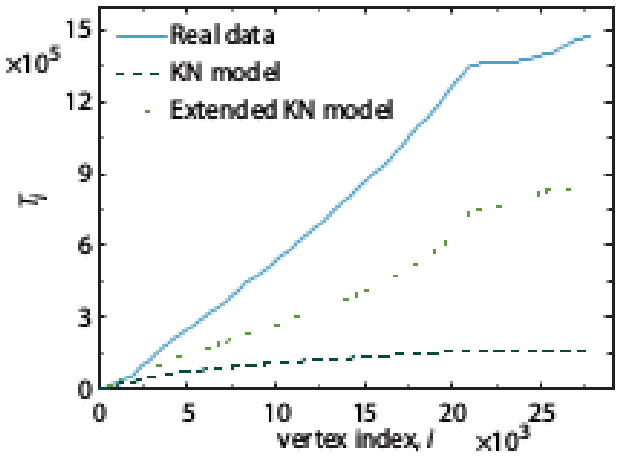

Figure 9: The graph shows the distribution of the total number of triangles in the network of 27 700 high energy physics papers comparing it with the simulation of KN model and the extended KN model. On the x axis it is reported the order of appearance, while T_i are all the triangles present in the network at the time of publication of paper i. Both models highly underestimate the real distribution

.

Both the KN model and the extension underestimate the number of directed triangles in the real network (fig 9).
So Wu and Holme propose a new model, very similar to the KN one, but more versatile in describing triangle formation.
'We start by ordering the vertices temporally as in the real data, and their out-degrees (the number of papers in the ref list) are kept the same as the original.
We do not restrict the number of in-degrees, that will be an emergent property we will use for validation.

A common assumption is that the relevance of a paper decays with its age: for this reason, we let the first arc from a new vertex i go to an old vertex with a probability proportional to its age $t_j = i - j$ to a power α (where a negative α reflects an attachment probability decaying with age).
To fill up the remaining out-degrees of i, we attach arcs with probability β to random (in- or out-) neighbors of j, and otherwise (i.e. with probability $1 - \beta$) attach arcs to older vertices with probability as above.
If there is no available neighbor to attach to (we assume one vertex cannot link to another vertex twice, or to itself), we make an attachment of the first type.In sum, our model has two input parameters α and β (in addition to the degrees), governing the two key ingredients, aging and triangle formation'.
In their analysis Wu and Holme keep as quantity of interest T_i, the number of triangles at the time of publication of i, and λ_i as defined above, the sum of in-degrees of the vertices that have been added in the network before i minus the sum of all in-degrees.
Modifying the values of α and β they obtain different distribution, that more or less resemble the real one (fig 10) .
From the comparison it is clear that the best results are given by the choice $\alpha = -1$ and $\beta = 0.99$, meaning that in the network the 99% of citations in the ref list of papers belong to the same cluster, while aging is described by an exponential with $\alpha = -1$.

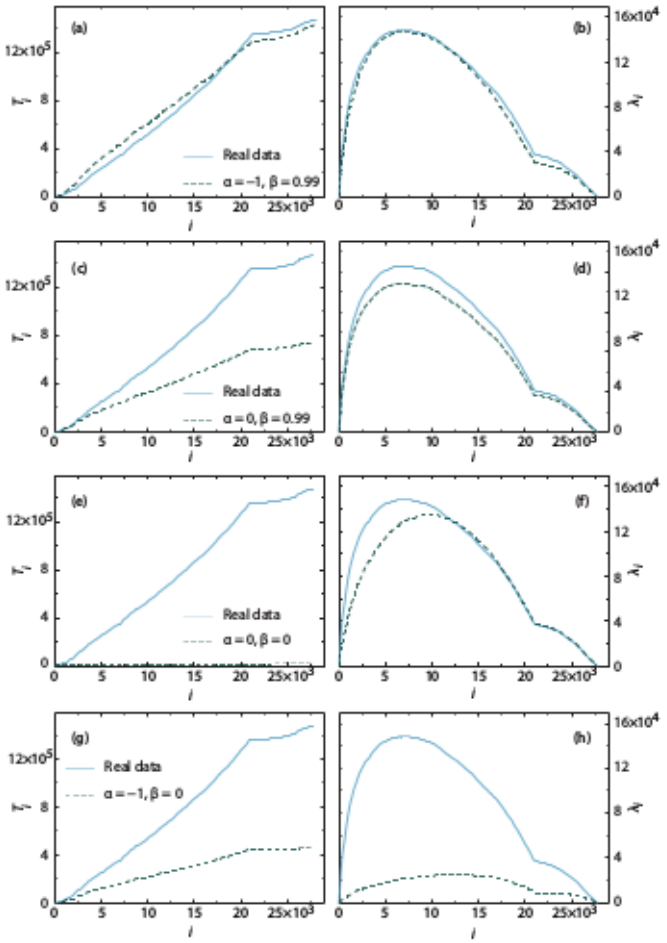

Figure 10: Solid lines correspond to real data, dashed lines to simulated networks with parameters specified in the box. As can be easily seen, the best results are given by the choice $\alpha = -1$ and $\beta = 0.99$

3.4 Redirection/Copying models

As we have seen in the previous sections, while *preferential attachment* accounts for the ubiquity of networks that are scale-free or have heavy-tailed degree distribution, it is too general and does not specifically address evolving network structures.
On the other hand, a more realistic scenario is provided by the two-step growth models that have been developed in the context of social networks and epidemic-like propagation of ideas.
All these models start considering that in writing a paper, an author reads research journals, searches the databases and finds some relevant papers, citing some of them in his reference list.
Then he studies the reference lists of these preselected papers, picks up relevant references, reads them, cites some of them, and continues this process recursively.
We will distinguish these two types of citation calling the first direct and the second indirect. Note that there is no *a priori* topological property that distinguish the first from the second ones.
This two steps process is constructed to account for the high number of triangles that is found in the data set of scientific citations: papers on the same topic very often have a lot of common citing papers.
A lot of models can be built with this mechanism, each one preserving different features: in the following I am reporting some of them.

3.4.1 Branching processes in citation dynamics

In 2003 paper [11] was published in which was stated that apparently, when an author forms the reference list for a paper, many of the articles he cites are simply copied from the ref lists of other papers: in an average reference list, only few papers are actually read.
Inspired by this discovery, in 2005 a first model was made by Simkin and Roychowdhury that took into account this copying mechanism, but on the other hand it did not take into account aging and preferential attachment.
Later on the idea was developed, and in 2007 the work 'A mathematical theory of citing' was published by the same authors [12].
In this model, referenced papers can be of two sorts

- Fresh papers the author has just read and uses in his work
- Older papers that have been cited by some recent paper

In the work Simkin and Roychowdhury consider as *recent* only papers published the preceding year, and in order to make the model mathematically tractable they use a time discretization with a unit of 1 year.
There are, on average, N_{ref} references in a published paper, and a fraction α goes to randomly selected preceding year works.
Obviously, if the number N of papers is large, the model leads to the first year citations being Poisson-distributed, and so the probability to get n citations is:

$$P(n) = \frac{\lambda_0^n}{n!}\, e^{-\lambda_0} \tag{6}$$

where $\lambda_0 = \alpha N_{ref}$, and in this context, citation dynamic becomes a branching process. To see in detail a bit more of this formalism, see appendix.
Using equation 6, we get the generating function for first year citations:

$$f_0(z) = e^{(z-1)\lambda_0}$$

and in general, if $\lambda = (1 - \alpha)$,

$$f(z) = e^{(z-1)\lambda}$$

As pointed out in the appendix, the process is much easier to analyze when $\lambda = \lambda_0$ or

$$\frac{\lambda}{\lambda_0} = \frac{\alpha}{1-\alpha} N_{ref} = 2$$

because this way all the generations of the branching process are governed by the same offspring probabilities.
If $P(n)$ is the probability distribution of the total number of citations a paper receives before being forgotten, we get

$$P(n) = \frac{1}{n!} \left[\frac{d^{n-1}}{d\omega^{n-1}} e^{n(\omega-1)\lambda} \right]_{\omega=0}$$

Using Stirling's formula, we get that for large n the distribution of citations is:

$$P(n) \propto \frac{e}{\lambda\sqrt{2\pi}} \frac{1}{n^{3/2}} e^{-(\lambda - 1 - ln\ \lambda)n} \tag{7}$$

When $1 - \lambda \ll 1$, the factor in the exponent can be approximated as:

$$\lambda - 1 - ln\ \lambda \sim \frac{(1-\lambda)^2}{2}$$

This way, for $n \ll 2/(1-\lambda)^2$ the exponent in equation 7 is approximately equal to 1, and $P(n)$ is dominated by the $1/n^{3/2}$ factor, while, when $n \gg$

$2/(1-\lambda)^2$, the behavior is dominated by the exponential.
Thus, we have a change in the behavior of the distribution for

$$n_c = \frac{1}{\lambda - 1 - ln\ \lambda}$$

However, if we try to analyze the real data, we get $\lambda \sim 0.1$, and this way we would obtain a cutoff at 200, too soon as the actual citation distribution obeys a power law well into thousands of citations.
This unwanted result can easily be solved including in our analysis the effects of literature growth and of *Darwinian fitness*.
Papers are not created equal, but each has a specific *fitness*, which is a measure of the scientific ability of the paper to 'fight' for citations with other competitors.
There can be different ways of defining such a property. In the paper in Ref [12] the authors consider a fitness bounded between 0 and 1, in such a way that a paper with fitness ϕ on average has

$$\lambda_0(\phi) = \alpha N_{ref} \frac{\phi}{\langle\phi\rangle_p}$$

first year citations, where $\langle\phi\rangle_p$ is the average fitness of published papers.
It is important to note that fitness distribution of references is different from the fitness distribution of published papers, as papers with higher fitness are cited more often. This distribution assumes an asymptotic form $P_r(\phi)$, which depends on the distribution of the fitness of published papers, $P_p(\phi)$.
With this conventions, during later years there will be on average

$$\lambda(\phi) = (1-\alpha)\frac{\phi}{\langle\phi\rangle_r}$$

next year citations per one current year citation for a paper with fitness ϕ, and $\langle\phi\rangle_r$ is the average fitness of a reference.
Now that we have introduced these concepts, calculating the average number of citations a paper with fitness ϕ acquires during its cited lifetime is relatively simple:

$$N(\phi) = \lambda_0(\phi)\sum_{n=0}^{\infty}(\lambda(\phi))^n = \frac{\lambda_0(\phi)}{1-\lambda(\phi)}$$

and so

$$N(\phi) = \alpha N_{ref}\frac{\phi}{\langle\phi\rangle_p}\frac{1}{1-(1-\alpha)\phi/\langle\phi\rangle_r}$$

and obviously $\langle\phi\rangle_r$ is obtained self consistently by averaging $\phi N(\phi)$ over ϕ:

$$\langle\phi\rangle_r = \frac{\int P_p(\phi)\phi N(\phi)d\phi}{\int P_p(\phi)N(\phi)d\phi}$$

If we consider as fitness distribution $P_p(\phi)$ a uniform distribution, the above equation becomes:

$$\langle\phi\rangle_r = \frac{\int_0^1 \frac{\phi^2 d\phi}{1-\gamma\phi/\langle\phi\rangle_r}}{\int_0^1 \frac{\phi d\phi}{1-\gamma\phi/\langle\phi\rangle_r}}$$

where

$$\gamma = 1 - \alpha$$

After some transformations this calculus reduces to

$$\gamma - 1 = \frac{(\gamma/\langle\phi\rangle_r)^2/2}{ln(1-\gamma/\langle\phi\rangle_r) + \gamma/\langle\phi\rangle_r}$$

When γ is close to 1 $\langle\phi\rangle_r$ is very close to γ, and replacing it in the latter we get:

$$\frac{\gamma}{\langle\phi\rangle_r} = 1 - e^{-1/(2(1-\gamma))-1}$$

Replacing this result in the equations above,considering λ as a function of ϕ and $\alpha = 0.1$, we get that the exponential cutoff for the fittest papers starts at about 300 000 citations. Moreover if we want to compute the overall probability distribution we need to average $P(n, \phi)$ over the (uniform) fitness distribution. After some tedious calculations we get

$$P(n) \propto \frac{e\langle\phi\rangle_r}{2\gamma}\frac{1}{n^2}$$

in the limit of $n \ll n_c$. In the opposite case, i.e. when $n \gg n_c$ we get:

$$P(n) \propto \frac{e\langle\phi\rangle_r}{2\gamma}\frac{\sqrt{n_c}}{n^{2.5}}e^{-n/n_c}$$

So, compared to the model without fitness, we have a modified power law exponent (2 instead of 2.5 for $n \ll n_c$) and a very much relaxed cutoff of this power law.
The major results obtained for the uniform distribution of fitness also hold for a non uniform distribution which approaches some finite value at its upper extreme $p_p(\phi = 1) = a > 0$. A wide class of fitness distributions produces results very similar to the ones just described.
Finally, we would like to include also the effect of literature growth: it is very well known and documented that the number of scientific publications grows every year exponentially, with an yearly percentage increase β that between 1970 and 2000 was ~ 0.045.Taking this effect into account, we get that

$$\lambda_0 = \alpha(1+\beta)N_{ref}\phi/\langle\phi\rangle_p$$

and

$$\lambda(\phi) = (1+\alpha)(1+\beta)\phi/\langle\phi\rangle_r$$

Obviously, $\langle\phi\rangle_p$ does not change with the introduction of β, since its only result is to increase the number of citations to all papers by a factor $1+\beta$, but while $\lambda(\phi)$ is always less than unity in the case with no literature growth, this is no longer true when we take this growth into account: when β is large enough, some papers can become super critical. The critical threshold at which this papers start to appear can be easily calculated as:

$$\beta_c = \frac{\langle\phi\rangle_r}{1-\alpha} - 1$$

Note that being in the super critical regime only means having extinction probability less than 1. With the uniform distribution of fitness we get that 1 in 400 papers become forever cited, but changing the distribution this fraction can be made much smaller.

3.4.2 Mean field approach

The model described above is a very good null model for redirection/copying models.
However it can not be considered as complete, since many of its basic assumptions are way too strong, e.g. direct citations only to precedent year papers.
Recently a new publication (Ref [13]) has appeared that, using the idea of separating citation dynamics as in Simkin's model, tries to be a bit more phenomenological.
First of all, for every paper it considers the function

$$R_0(t_0) = \int_0^\infty R(t_0, t_0 - t)dt$$

that describes how the number of papers in reference lists varies with time.
Let t_0 be the publication year of the paper and $R(t_0, t_0 - t)$ the number of papers in the ref list published in the year $t_0 - t$.
In the two step model we can separate direct and indirect references, $R_{dir}(t_0, t_0 - t)$ and $R_{indir}(t_0, t_0 - t)$, and the approximation that the authors consider is that once someone cites some paper, he can cite any of its references with equal probability. This way, an average reference list comprises $R(t_0, t_0 - \tau)$ preselected papers published in year $t_0 - \tau$, and the fraction of references in

year $t_0 - \tau$ directed to papers published in $t_0 - t$ (with $t < \tau$) is:

$$\frac{R(t_0 - \tau, t_0 - t)}{R_0(t_0 - \tau)}.$$

Since the age composition of the reference list is fairly independent of the publication year, we have that

$$\frac{R(t_0 - \tau, t_0 - t)}{R_0(t_0 - \tau)} = \frac{R(t_0, t_0 - t + \tau)}{R_0(t_0)}$$

The mechanism starts with direct citations, and continues randomly citing papers from their reference lists and iterating this procedure.
Anyway, even though the selection of papers follows a uniform distribution within a ref list, the number of indirect citations from a paper depends on time: $T(\tau)$ is the number of indirect papers that are added in the ref lists of papers published in t_0 coming from the copying of ref lists of papers published in $t_0 - \tau$. Said so, we have

$$R(t_0, t_0 - t) = R_{dir}(t_0, t_0 - t) + \int_0^t R(t_0, t_0 - \tau) \frac{T(\tau)}{R_0(t_0)} R(t_0, t_0 - (t - \tau)) d\tau \quad (8)$$

and the first term comes from direct references, while the second comes from the procedure described above, and accounts for indirect citations.
Thanks to the fact that there is an obvious duality between references and citations, the age distribution of references $R(t)$ is very similar to $M(t)$, the mean citation rate of papers published in one year.
Consider a set of all $N_0(t_0)$ papers in a certain research field published in year t_0. The mean number of citations that a certain paper garners in the t-th year after publication is $M(t_0, t_0 + t)$, and this should be equal to the mean total number of references in year $t_0 + t$, so that

$$N_0(t_0) M(t_0, t_0 + t) = N_0(t_0 + t) R(t_0 + t, t_0) \quad (9)$$

This equation implicitly takes into account that both the number of publications N_0 and the reference list length R_0 grow exponentially with time:

$$N_0(t_0) \propto e^{\alpha t_0}$$

$$R_0(t_0) \propto e^{\beta t_0}$$

We arrive to a mathematical expression for the reference-citations duality:

$$M(t_0, t_0 + t) = e^{(\alpha + \beta) t} R(t_0, t_0 - t)$$

Substituting into eq 8, we get a dynamic equation for the mean citation rate:

$$M(t) = M_{dir}(t) + \int_0^t M(t-\tau)\frac{T(t-\tau)}{R_0}M(\tau)d\tau$$

where, since all the quantities depend on t_0, I've kept only the t dependence, meaning that $M(t) = M(t_0, t_0 + t)$, and $M_{dir} = R_{dir}(t)e^{(\alpha+\beta)t}$.
$T(\tau)$ has been replaced by $T(t-\tau)$ using the properties of the convolution, and $\frac{T(t-\tau)}{R_0}$ is the probability of indirect citation at time t via another paper published in year $t - \tau$.
From this equation, after calculating from data $M_{dir}(t)$, the authors found an exponential kernel $T(t) = T_0\ e^{-\gamma(t)}$, with $T_0 = 6.6$ and $\gamma = 0.64yr^{-1}$.
Said so, finding the age distribution of references is an easy task from eq 9, knowing that the growing in number of publications has parameter $\alpha = 0.046$, while the growth in reference lists length parameter is $\beta = 0.02$.

3.5 Describing citations of single papers: lognormal aging

Another type of models starts from considering the history of each single paper.
If $\Pi_i(\delta t_i)$ is the probability that a paper i is cited by another paper at time t_i after publication, we can, for example, separate the contributions of preferential attachment, of fitness and of aging, considering them independent. This way, we get:

$$\Pi_i(\Delta t_i) \sim \eta_i\ c_i^t\ P(\Delta t_i) \tag{10}$$

where η_i is the fitness of the paper, c_i^t the number of citations it has achieved at time t, and $P(\delta t_i)$ the temporal relaxation function, that describes the tendency of papers to diminish the rate of citations with the passing of time. This last function, in principle, can be computed directly from the data set. However, in order to do so we should group papers with the same fitness (η) and cumulative citations (c_t), and look at the time when they are cited again. There are technical problems in doing this: η is a very difficult value to get, and normally aging is very dependent on the topic, i.e. fields with higher number of researchers normally have higher number of publication every year, and consequently a much higher citing rate.
In the paper in Ref [14] a formula is calculated for the description of cumulative distribution of citations which considers $P(\Delta t_i)$ a log-normal distribution:

$$P(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma\Delta t}exp(-\frac{(log\Delta t - \mu)^2}{2\sigma^2})$$

Lognormal aging is a very common tendency of many real systems: it provides an excellent fit in a lot of phenomena involving reaction times (RT) in biology, economics, and other branches of population dynamics.
In the paper 'Information processing models generating lognormally distributed reaction times', by Ulrich and Miller (Ref [15]) the authors somehow motivate why this distribution is so common among these phenomena.
In general, a random variable T follows a lognormal distribution if $log(T)$ follows a normal distribution with mean μ and standard deviation σ (T must be a positive random variable).
Two processes have been discovered to generate it. The first arises as a transform of a fluctuation of a random normal noise, while the second leads to the lognormal directly, without any other transform.
In the *Logarithmic activation growth* a stimulus that requires a speeded response is presented at $t = 0$. Let $A(t)$ be the response activation function that begins to accumulate as a logarithmic function of t:

$$A(t) = k\ ln(t)$$

The response is triggered when A(t) reaches a certain threshold $C > 0$, but C is subject to random trial to trial fluctuations. If T is the instant when $A(t)$ reaches the C level, we have:

$$A(T) = C,$$

$$k\ ln(T) = C$$

and so

$$T = exp(T/k)$$

T will be lognormally distributed if C follows a normal distribution.

The second is called *partial output model*. Starting from the late 70s many theorists have considered models in which RTs are determined by a series of different processes that trigger the activation from an input to an output level.
Partial output models can be described as a composition of n successive processes, of which the output from the $(i-1)$ unit serves as the input of the i-th.
This way, the output of the n-th unit is described by the function:

$$O_n(t) = g_n(...g_3(g_2(g_1(t)))...)$$

where $g_i(t)$ is a monotonic increasing function denoting the output of unit i at time t.

The lognormal shape arises if $g_i(t)$ is a power function:

$$g_i(t) = A_i[g_{i-1}(t)]^{b_i}$$

where A_i is a positive random variable and b_i a positive constant.
The first unit gets the value:

$$g_1(t) = A_1(t - t_0)^{b_1}$$

if $t > t_0$, otherwise for $t < t_0$ it is zero. Finally, the response signal is triggered if $O_n(t)$ reaches a constant level $c > 0$.
When n is large enough, under the general assumption of the central limit theorem, we get that T, the random variable that describes the phenomenon $O_n(t) = c$, is lognormally distributed (shifted by t_0).

Once we have justified the choice of the lognormal aging distribution,with equation 10, we start considering that, if c_i^t is the number of citations a paper i has at time t, considering eq 10, we have:

$$\frac{dc_i^t}{dN} = \frac{\Pi_i}{\Sigma_{i=1}^N \Pi_i}$$

In considering $N(t)$, the total number of published papers, we have that

$$N(t) \propto exp(\beta t)$$

where $\beta = (17year)^{-1}$ for the PR corpus. This way, to transform $\Delta t_i = t - t_i$ in something related to the mean number of papers published since time t_i of publication of paper i, we have

$$\Delta t_i = \beta^{-1} ln(N/i)$$

so that:

$$\frac{dc_i^t}{dN} = m \frac{c_i \eta_i P_t(\beta^{-1} ln(N/i))}{\Sigma_{i=1}^N c_i \eta_i P_t(\beta^{-1} ln(N/i))}$$

and assuming $c_i = m(f(\eta_i, \Delta t_i) - 1)$, after some tedious calculation we find:

$$c_i^t = m \left(e^{\frac{\beta}{A} \eta_i \Phi\left(\frac{ln(t) - \mu_i}{\sigma_i}\right)} - 1 \right) \tag{11}$$

where $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2} dy$$

and as β and A are system parameters, we will use $\lambda_i = \beta\eta_i/A$ as the general form of the fitness of paper i.
Particularly interesting of this formula is that the ultimate impact of a paper, i.e. the total number of citations it gets, is achieved via a remarkably simple formula: since in the limit $t \to \infty$ the function $\Phi \to 1$,

$$c_i^{\infty} = m(e^{\lambda_i} - 1)$$

and so this quantity depends only on the fitness of the paper and not on any other of the parameters we have considered.
The only true weakness in this is that for every paper in eq 11 we have three fitting parameters, and this does not allow us to make real predictions on the evolution of its citing life within reasonable mistakes.

Having considered these as the most representative of the models of citation dynamics, we immediately notice that in all of them the emphasis is put on time distribution and on triangles. However, as we will see, these two ingredients alone are not sufficient to describe many of the most important topological features of the network.

4 Sleeping Beauties and delayed impact papers

Normally, when a paper is published, it gets immediately a lot of attention and reaches the peak of annual citations in the first two or three years.
But this is not always the case: sometimes it happens that a paper does not get almost any attention for many years, even for decades, until, at a certain time, it experiences an 'explosion' of citations.
'Being ahead of one's time' is something that has always intrigued and frightened scientists, but the first systematic studies on this subject started only in 2004, with the paper 'Sleeping Beauties in Science' by Van Raan [16].
'Some scientists claim that some of their publications have not been as successful as they should because they are ahead of time'. We call this the 'Mendel syndrome', named after Gregor Mendel, whose discoveries in plant genetics were so unprecedented that it took thirty-four years for the scientific community to catch up to it.
As the number of citations of a paper is taken by scientists as a proxy to the importance of a paper, understanding why sometimes this reckoning is delayed is a very important target.
We will call 'Sleeping Beauties' (SB in the following) papers that, after publication, experience a period of 'sleep', i.e. many years in which they do not get as many citations as they should, and that, at a certain point, start gaining a lot of them.
In his work, Van Raan considers three main variables as thresholds to distinguish between SB papers and non SBs:

- *depth of sleep*: a deep sleep has at most 1 citation per year, while a less deep sleep has between 1 and 2 citations per year
- *length of sleep*: length of the period of sleep, as described above
- *awake intensity*: number of citations per year, for four years after 'awakening'. The rate of publications in this period has to be big, much bigger than during the sleep.

Using a very large data system, with about 20 million articles, Van Raan took papers with 6 different sleeping periods, s=5,6,7,8,9,10 years, all published in 1980.
He identified the articles in deep sleep or in less deep sleep as mentioned above, while for the awake intensity he defined 5 different classes, grouping papers with $c_w = [21, 30], [31, 40], [41, 50], [51, 60]$and $[> 60]$ citations, i.e., on average 6, 9, 12, 15, and more than 15 citations per year during the four-year

awakening period.

What Van Raan found is summarized in the table of fig 11:

c_w	N, less deep	N, deep
[21,30]	276	41
[31,40]	29	6
[41-50]	5	0
[51-60]	0	1
[> 60]	1	0

Figure 11: Table that summarizes the SBs found by Van Raan in his analysis. With his definition, even with a very big data set, we get very few SBs

In the end, in his analysis he concludes that:

- The probability of awakening after a deep sleep is smaller for longer sleeping periods
- long sleeping periods are less likely for less deep sleeps
- The awakening intensities is independent of both depth and length of sleep!

4.1 Awakening of a SB: the Prince

After this first systematic analysis of the delayed impact phenomenon a lot of effort has been put to get as much information as possible on it.
In particular, very interesting is the mechanism that triggers the awakening: why, all of a sudden, a paper starts getting citations at a much higher rate than before?
A lot of hypothesis have been made on this, and many scientists have started to think that, around the awakening year, SBs are cited by a very good paper, that giving new visibility to it brings a whole new set of citations.
As a consequence of Van Raan's appealing metaphor of 'Sleeping Beauties', this new paper was called 'The Prince'.
Van Raan himself introduces this idea in his work, naming the Prince (Polchinski, 1995) of a Sleeping Beauty (Romans, 1986), but except for this single example, no thoughts were given to the Prince's role in more general terms.
In 2010 Braun et alii published a paper called 'On Sleeping Beauties, Princes and other tales of citation distributions ...' [17] in which they try to fill this gap by performing a detailed citation analysis of a set of Sleeping Beauties together with that of the corresponding Princes.
In their work it is suggested that the same mechanism ('induced citations', i.e. attention given to a paper after being cited by a later and more visible one) may also work for 'normal' (non-delayed) citation histories.
Braun took papers of 1980 from 'science citation index' that were not cited in a period of 3 to 5 years after publication. Then he searched for the prince among the citations in their sleeping period: 'candidate Princes were sought for among the first citing articles; they were supposed to be highly or at least fairly cited and to have a considerable number of co-citations with the Sleeping Beauty'.
Anyway, this idea of a 'Prince' immediately encounters some problems: in the example above, mentioned by Van Raan, we have that the SB got 256 citations and the prince 1225, but the number of co-citations is only 56: a very small fraction of the total.
Braun sample covered a wide range of scientific fields. Another interesting discovery was that almost the 40% of what he identified as princes came from other disciplines: in the table below are summarized SBs and princes grouped by subject (Fig 12).

Very often, princes are papers that are published in journals of higher impact factors: the average impact factor of the journals in which the PRs were published is more than twice as high as that of the SBs.
Another very important remark is that, after the kiss, both the prince and the SB have a very long and successful life, and even though there is a sub-

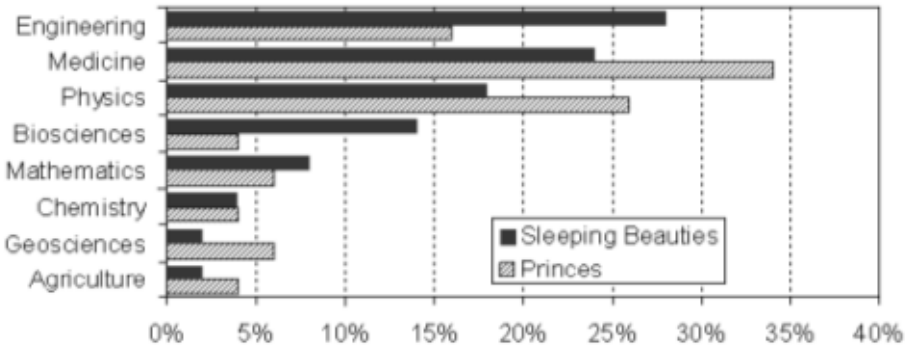

Figure 12: Percentage of SBs and princes by research field. In the 40% of the cases, prince and SB do not belong to the same field

.

stantial part of common citations, their lives are independent.
So the idea of the 'Prince' that wakes the sleeping article is very interesting, but it seems to be too naive in the cases analyzed: a more systematic approach is required.

5 SB coefficient

The ideas introduced by Van Raan to treat delayed recognition have been very useful.
However, as we have already mentioned, the definition he gives, together with the strict thresholds he puts, produce a relative scarcity of SBs that does not allow to make any significant statistical analysis.
On the other hand, some delayed impact papers (recognizable by looking at their cumulative citation distribution) are not included in that first definition.
For example, a very clear defect is that it does not consider what happens after the awakening in putting thresholds on what happens before: for papers with thousands total citations experiencing a period of 10 years with only 10 or 20 citations means delayed recognition, even if they have not been 'sleeping' in Van Raan's definition.
The cumulative distribution of citations very often has the shape drawn in fig 13, with no change of concavity.

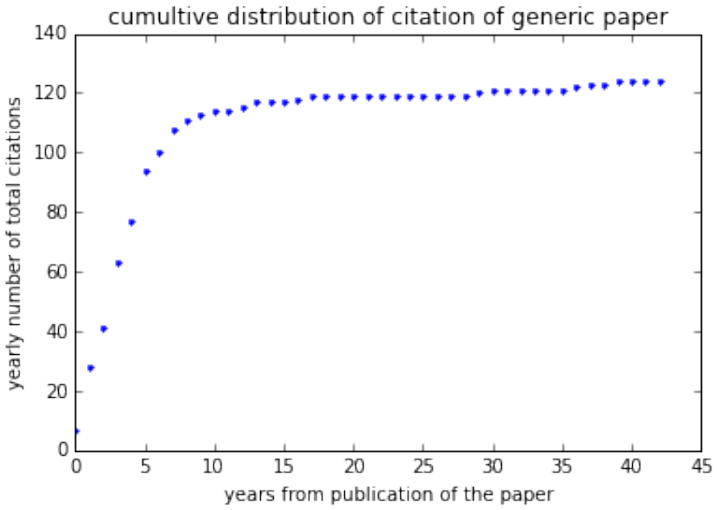

Figure 13: Cumulative distribution of citations of a normal paper:the distribution has no concavity change

.

However many other distributions do not have such a simple shape: in fig 14 there is an example of a more complicated distribution. Note that, in Van Raan's definition, this is not considered as a SB.

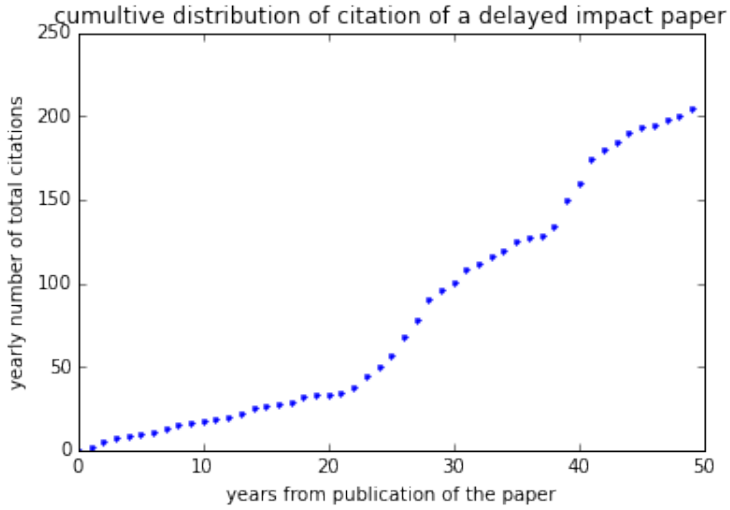

Figure 14: Cumulative distribution for a delayed impact paper: it does not experience any period of stop, but nonetheless its citing life has a delayed impact

In the first months my work was focused on changing this first definition of SB.
While I was doing so, the paper *Defining and identifying Sleeping Beauties in science* [18] was published, by Qing Ke, Filippo Radicchi et alii, that tries to do, more or less, the same thing.
In their analysis, the authors propose an index, called the *Beauty coefficient*, denoted as B, that can be calculated for any given paper and is based on the comparison between its citation history and a reference line, drawn from its publication year to the year of the peak of citations.
Let's call t the time interval after publication and c_t the citation history of the paper, i.e. for every t c_t tells how many citations the paper got in the t-th year after publication.
If c_{t_m} is the maximum of c_t, with $t_m \in [0, T]$, the straight line l_t that connects the point $(0, c_0)$ and (t_m, c_{t_m}) analytically is described by the equation

$$l_t = \frac{c_{t_m} - c_0}{t_m} t + c_0$$

For each $t < t_m$ we can compute the ratio between $l_t - c_t$ and $max\{1, c_t\}$,

and the definition of B is achieved by summing over these values (Fig 15):

$$B = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m}-c_0}{t_m}\ t + c_0 - c_t}{max\{1, c_t\}}$$

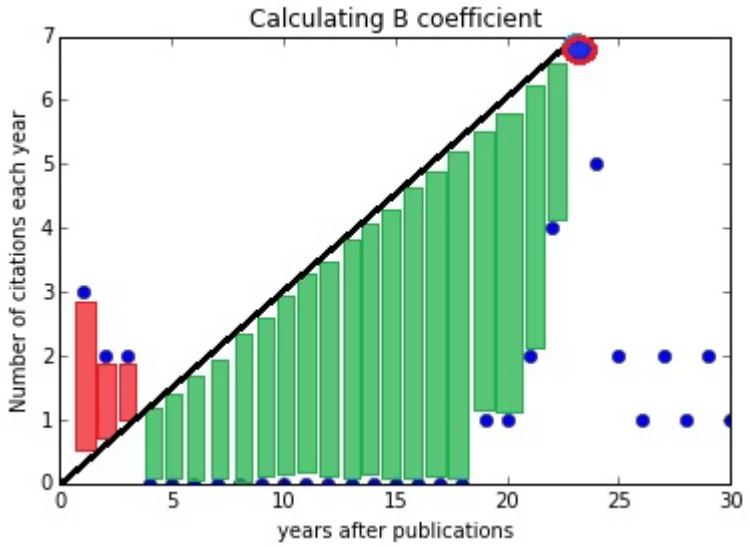

Figure 15: The blue point with the red border is (t_m, c_{t_m}),i.e. the maximum. The line l_t connects it to $(0, c_0)$. For each t after publication we can calculate the area of the bin between the point and the line: if $c_t > l_t$ the area is considered negative (red rectangles), while if $l_t > c_t$ it is considered positive. B coefficient is found summing each bin's area divided for the number of citations achieved in that year ($max\{1, c_t\}$ is to avoid division by zero)

Some features of this definition:

- for papers with $t_m = 0$ the maximum is immediately achieved and $B = 0$. This is a very common situation
- papers with c_t growing linearly have $B = 0$
- Van Raan's definition of delayed impact strongly depends on the choice of thresholds, while B does not have any
- B increases both with the length of sleep and the awakening intensity

What is strange of this definition is the high importance given to the peak: to legitimate this choice, the authors state that for most of the papers

yearly citation count decreases very quickly after reaching it.
The advantage of this definition for SB is that it does not rely on arbitrary thresholds. This way, we can investigate this phenomenon at a systematic level.
What is observed is a heterogeneous but continuous distribution of B, and most of all no clear demarcation that separates SB behavior from normal one (fig 16).

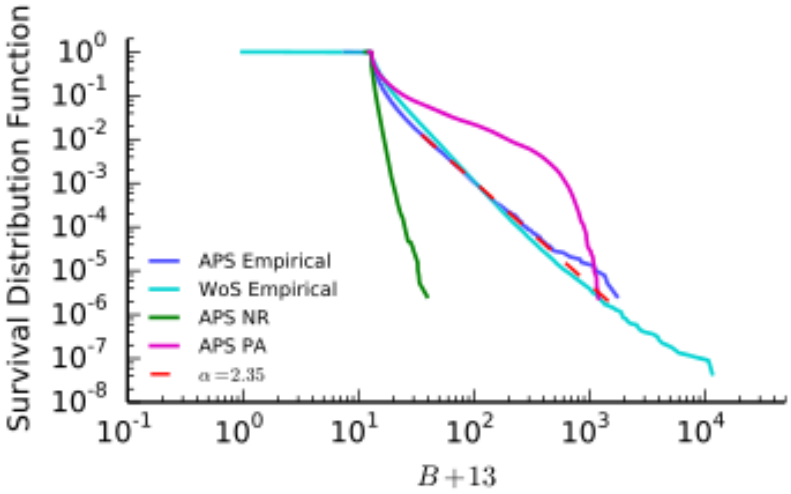

Figure 16: Distribution of the B coefficient. Since B values can be negative, x axis is shifted by 13 in order to put the logarithmic scale on it. The blue and cyan curves represent the empirical results, and the ones from NR and PA model are plotted as green and magenta lines. The red dashed line is the best estimate of a power-law fit

.

After giving this definition, Ke et alii compare the distribution of the beauty coefficient with the one predicted by two null models. The first one is the citation network randomization (NR), in which, constraining both the number of papers in the reference list and the number of citations, we swap the links between them (very similar to the model proposed by Karrer Newman, described in section 3.3).
The other null model is preferential attachment (PA), already explained in section 3.1 (note that with it the definition of B is not even consistent, since there is no drop in citation count after the peak).
The result, presented in Fig 16, is that neither of them satisfyingly describes the real behavior.
Moreover, the occurrence of SBs is not a phenomenon that can completely

be understood just by looking at single subject citation histories.
Fig 17 shows the contribution to the total number of SBs divided by discipline: the first three are multidisciplinary subjects, and together bring about the 23% of the total. So interdisciplinarity seems to play a huge role in what triggers the awakening of papers.

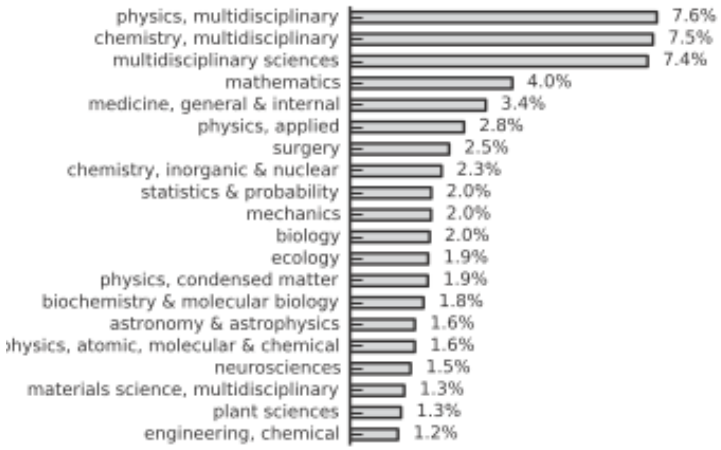

Figure 17: This is the top chart of disciplines producing SBs as found in the paper by Ke et alii [18]. The first three are multidisciplinary areas.

.

In order to be more quantitative and analyze the contribution of these external citations, Ke at alii decide to separate the set of SBs into three disjoint subsets with high, medium and low values of B. For each paper they compute the cumulative distribution of external citations.
From Fig 18 one can conclude that top SBs are clearly very influenced from external subjects, with about an 80% of papers that have a 75% interdisciplinary citations.

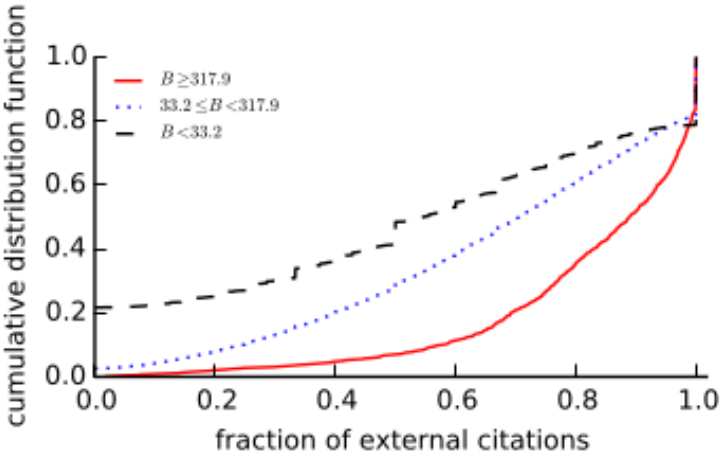

Figure 18: The graph shows how interdisciplinarity affects the behavior of SBs. The image represents the cumulative distribution of the fraction of external citations, grouped in 3 subsets of low, intermediate and high B values

.

5.1 Modifying B algorithm

The algorithm just introduced has been very useful to prove that *delayed impact* is not a pathological property of a separated set of papers.
However when I tried to use this definition I encountered some problems: papers that could undoubtedly be considered SBs where not recognized as such.
What I found to be the main weaknesses of Ke's B coefficient are:

- it is binning dependent: if instead of taking time skip one year one takes two years or six months, the value changes abruptly (see figures 19 and 21)
- it does not take into account that papers' citation histories have different time scales: for many papers it is unusual to have more than one citation per year, but the frequency of their occurrence may determine an actual delayed recognition

B coefficient works really well with top class SBs, that after discovery have huge numbers of citations every year, but for lower level SBs (i.e. up to 50 total citations) it gives some unwanted results: many of the papers I had found as SBs had very low B values.

Fig 19 and 20 report the year citation distribution and cumulative distribution of a paper: Bc is zero, because the peak of citations is achieved in the first year, but from Fig 20 one still concludes that it is indeed a delayed impact paper!

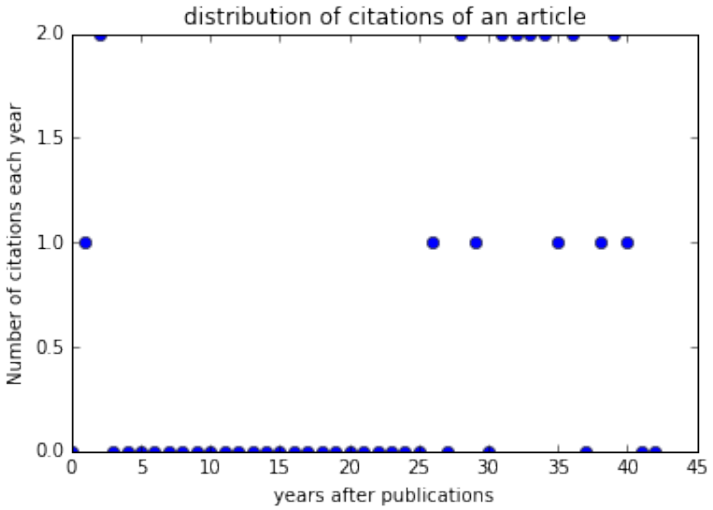

Figure 19: X axis are the years after publication, y axis reports the number of citation per year. The peak is in year 2, and so $Bc = 0$

.

I've tried to modify B coefficient in order to avoid this unwanted results. An estimate of the SB behavior should manifest some regularity properties, such as:

- show dependence on the citation life of the paper after the awakening, growing both with the delay and the total amount of citations
- not to decrease if the number of citations grows
- not strongly depend on just one year's citation history

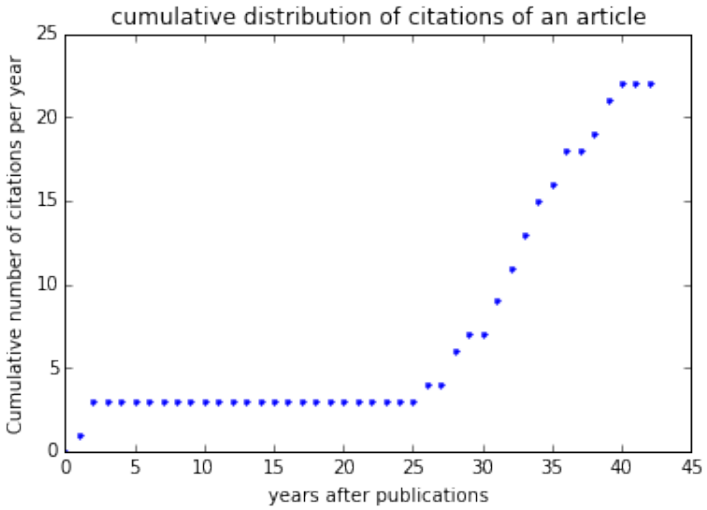

Figure 20: Although $B = 0$, the paper can be easily recognized as a SB from its cumulative distribution

.

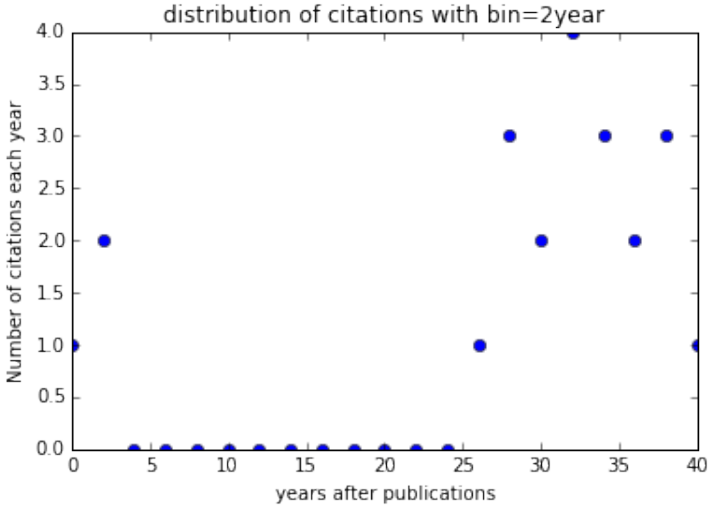

Figure 21: Distribution of citations as in Fig 19 but with binning 2 years instead of 1. Since the peak is now not at the beginning but around the 33^{rd} year after publication, B coefficient is not zero (~ 29). This manifests some problems in the consistency of the coefficient.

In order to fulfill these requirements, I've used the cumulative distribution of citation and not just the normal one.
The new SB coefficient can be defined as follows.

If $y(\Delta t)$ is the cumulative distribution of a paper and Δt the time from its publication, we consider for every Δt the line from the origin of the graph (the point $(0,0)$) to that point, and compute the area between this line and $y(\Delta t)$. We take as 'SB coefficient' the biggest of these areas. More precisely,

$$m(\Delta t) = \frac{y(\Delta t)}{\Delta t}$$

is the slope of the line from $(0,0)$ to point $(\Delta t, y(\Delta t))$. For every Δt (zero excluded) we can calculate the area (Fig 22 and 23)

$$SB(\Delta t) = \sum_{\Delta t'=0}^{\Delta t'=\Delta t} [m(\Delta t)\Delta t' - y(\Delta t')]$$

and we will call 'SB coefficient' the highest of these values

$$SBc = max_{\Delta t} SB(\Delta t).$$

and ΔT the time that maximizes it.

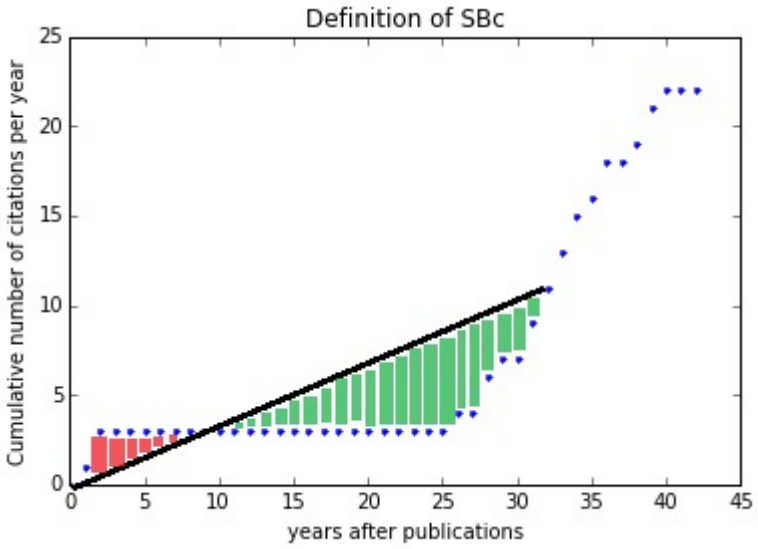

Figure 22: Once we have chosen a point of the graph, we can draw a line from the origin to it and consider the area (with sign) between the distribution of citations and it. In this case, the bins in green contribute positively, while red ones negatively. We calculate this area for every point of the graph, and the highest value is the SBc

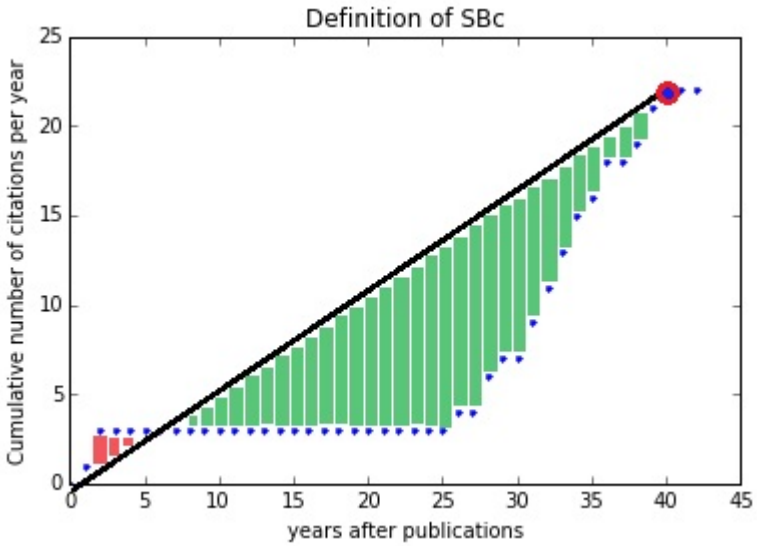

Figure 23: The graph shows the configuration of the line that gives the biggest area. The Δt of the point with the red border (the one that maximizes the area) is called ΔT

With the new definition for the paper with zero B value of figures 19 and 20 we get $SBc = 188$. As already pointed out in Van Raan's paper, there are a lot of different features with which one describes delayed recognition, and SBc is far from describing alone all of these: different papers with different citation lives might have equal SBc.
So in order to make comparisons (and put thresholds) we need to define other coefficients.

5.1.1 Awakening year

Once we have fount ΔT and the line from the origin to $(\Delta T, y(\Delta T))$, we can consider distances of points of $y(\Delta t)$ from it. Borrowing the definition given in the paper [18] (but slightly modifying it), the awakening time (awt)is defined as the year in which $y(\Delta t))$ is the farthest from the line (Fig 24). Formally, if $d(\Delta t)$ is the distance of point $(\Delta t, y(\Delta t))$ from the reference line, we have:

$$d(\Delta t) = \frac{y(\Delta T)\ \Delta t - \Delta T\ y(\Delta t)}{\sqrt{y(\Delta T)^2 + \Delta T^2}}$$

and so

$$awt = \{\Delta t : d(\Delta t)\ \text{ is maximized}\}$$

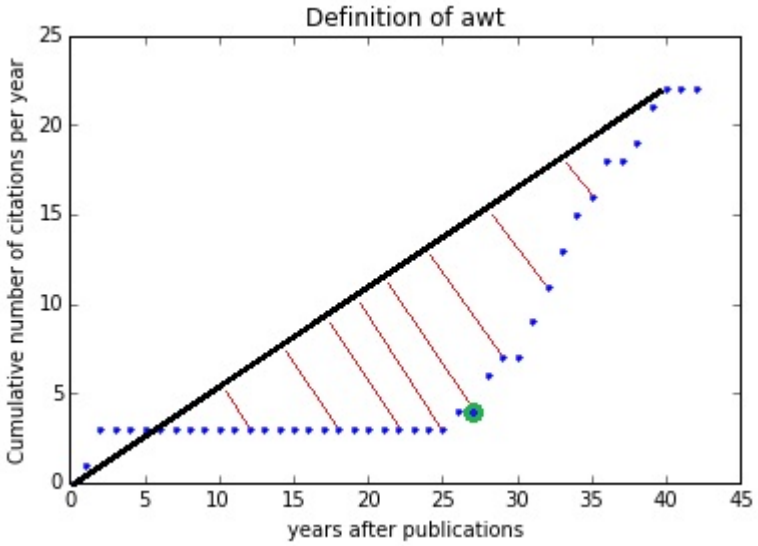

Figure 24: For each point we calculate the distance from the reference line (i.e. the line that maximizes the area in the SBc definition). The Δt point that stands at the longest distance from it is the awt of the paper. In this graph, the farthest point is the one with the green border, and its year is the awt (so in this case $awt = 27$).

5.1.2 Depth of sleep

SBc depends both on what happens before and after the awt, but very often, in order to compare how much attention a paper has achieved before awt, one needs another coefficient.
The *Depth of sleep* of a paper (*dos* in the following) is the ratio between the area between the reference line (from the origin to $(\Delta T, y(\Delta T))$) and $y(\Delta t)$ (with sign) and all the area under the line, both calculated only until awt (Fig 25).
If $m(\Delta T)$ is the slope of the reference line:

$$dos = \frac{\sum_{\Delta t=0}^{\Delta t=awt}(m(\Delta T)\Delta t - y(\Delta t))}{0.5\ m(\Delta T)\ (awt)^2}$$

Note that *dos* has 1 as maximum value, but can also have negative values. 'Second life papers' are the ones that were awake in the past, and that experienced a new period of sleep.

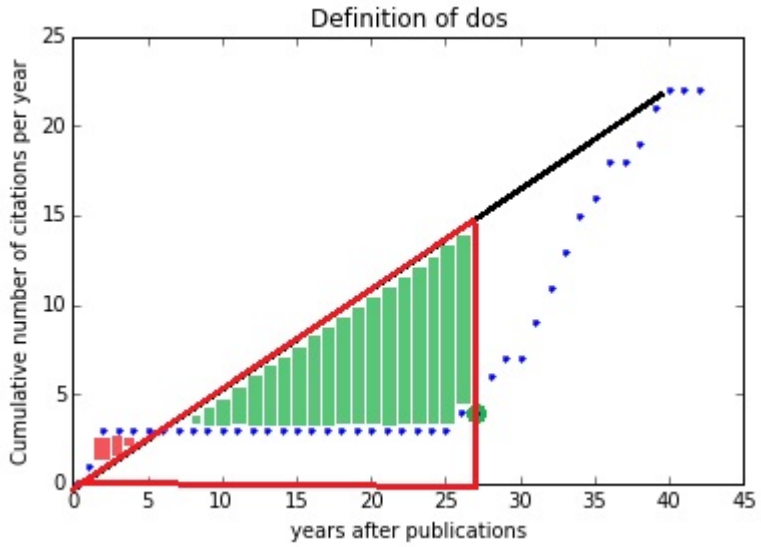

Figure 25: keeping the line that maximizes SBc, we calculate the area (with sign, as before) between l_t and y_t, but now only until awt. Then we divide this value for the area of the triangle (the one with the red lines in the picture) of vertices $(awt, m(\Delta T)\ awt)$, $(awt, 0)$, $(0, 0)$. *dos* has 1 as maximum value, when the paper has not had any citations before awt, but can also be negative!

6 Statistics of SB using SBc, awt, dos

Once we have defined the coefficients above, we can analyze their distributions in the APS dataset.
About the 55 % of papers have zero SBc, so the 45% experience some sort of delayed impact. However only the 15 % have $SBc > 5$.
The first interesting quantity is the number of total citations SBs get compared to the one of normal papers.In the graph in Fig 26 I've compared them.

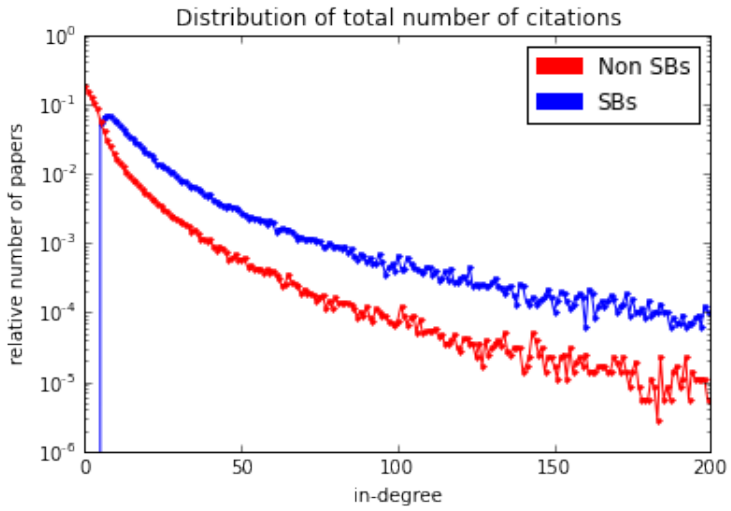

Figure 26: Blue dots are SBs, while red ones are normal papers. There is no SB with less than 4 citations

.

The difference between the two distributions may be a consequence of the fact that there is no SB with less than four citations. However, even if we consider only papers with more than ten citations, still delayed impact papers tend to have more citations than normal ones (fig 27).

In Fig 28 there is the distribution of SBc for papers in the APS dataset

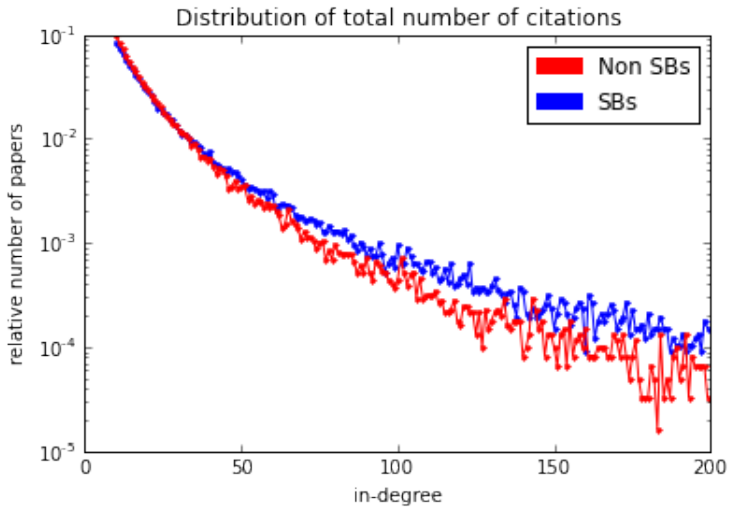

Figure 27: Distribution of papers with more than ten citations. The two lines are still substantially different.

.

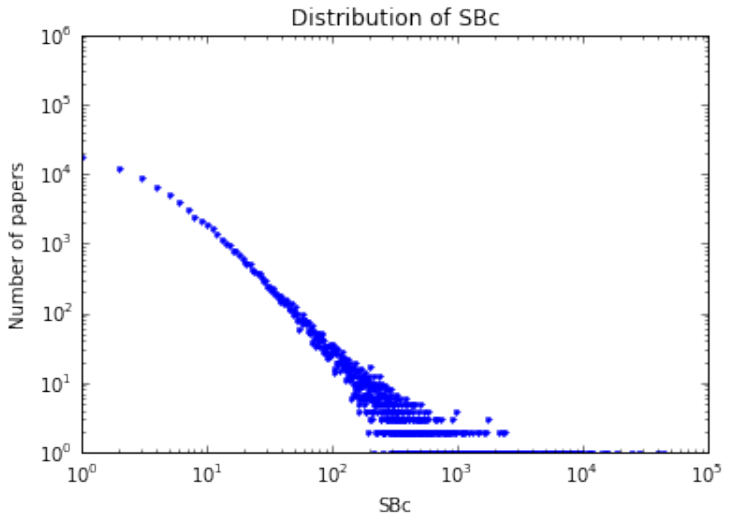

Figure 28: x axis are SBc values, while y values are the respective number of papers. The scale of both axes is logarithmic. The total number of papers with non zero SBc is about 202 000 over 450 000 studied papers, so about the 45%. The highest SBc is about 45 000

.

Together with the SBc, another interesting distribution is the awakening time (Fig 29).

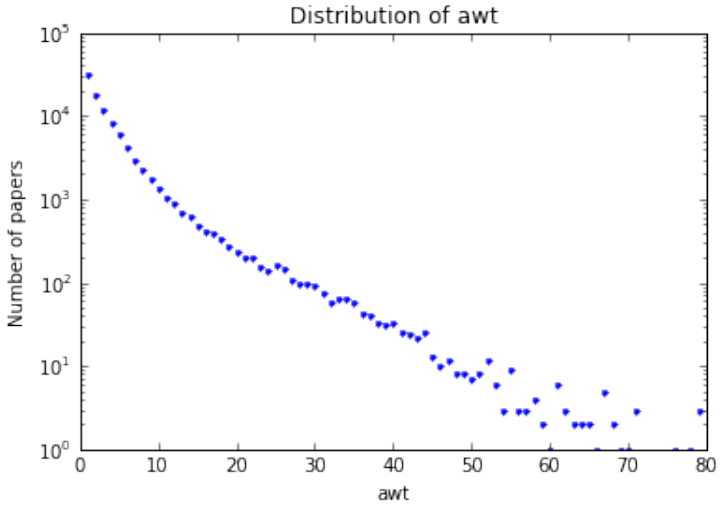

Figure 29: Data are reported only for non zero SBc (papers with $SBc = 0$ have awt=0). x axis are awt values, while on the other axis there is the respective number of papers. The scale of y is logarithmic. The distribution could be modeled as the sum of two exponentials.

.

The distribution of the ' depth of sleep' is not very significant, but, on the other hand, we can draw the dependence of the awt and dos in a two dimensional graph (Fig 30), and as already pointed out in the paper in [18], from it we can see that the distribution is not separated: there is no special behavior of some papers that experience delayed impact, but a continuous distribution that, starting from the peak $SBc = 0$, decreases regularly.

We might also want to compare SBc values of papers that had $B = 0$. In the APS dataset, I have found 12 000 such papers, and in fig 31 I have reported their SBc distribution. Note that there are very high SBc papers that are completely ignored using B coefficient.

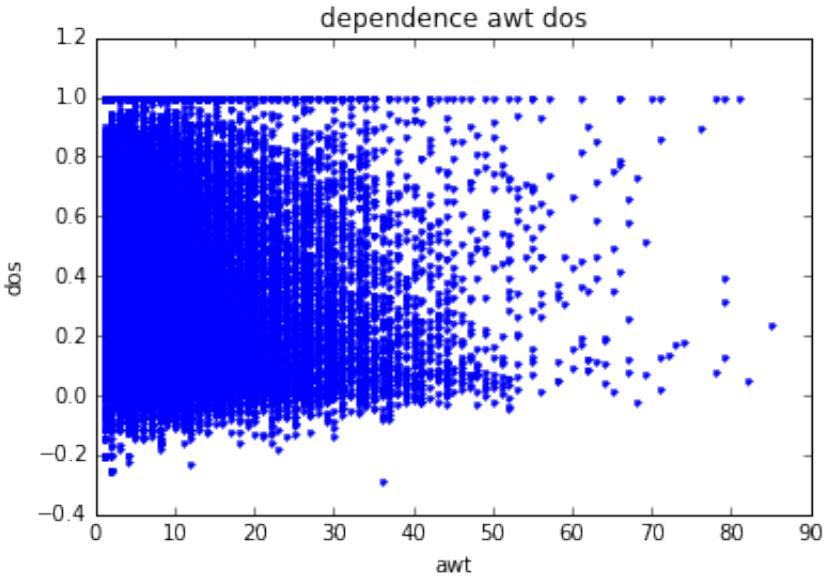

Figure 30: Dependence of the depth of sleep from the awakening time. As already mentioned, while *dos* values cannot be higher than 1, they can be negative

.

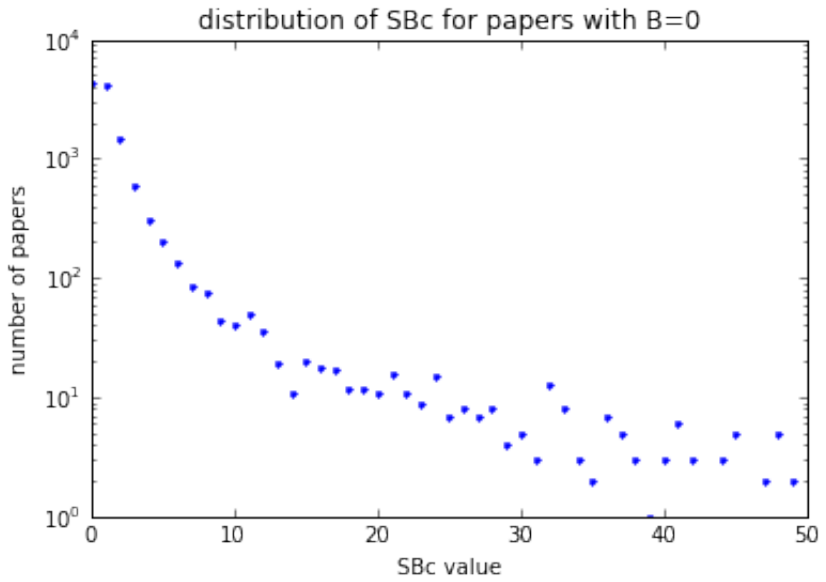

Figure 31: Distribution of the values of SBc for papers with zero B values

.

6.1 Closeness of SBs

Another important topological feature is that SBs very often tend to cite each other.
If we compare the number of SBs in the *reference lists* of normal papers with the one of SBs (fig 32), the two distribution are significantly different: the fraction of SBs in a SB reference list is on average much higher than the fraction for normal papers.

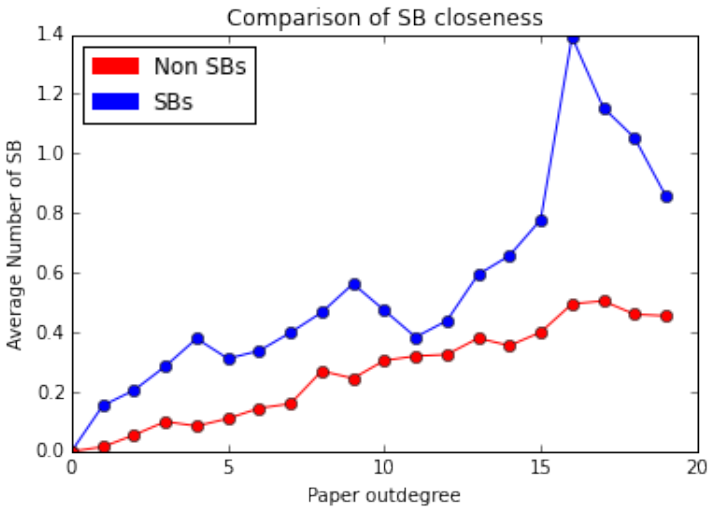

Figure 32: To draw this graph, I've taken all the SBs with $awt > 7$ and $dos > 0.7$, put together the ones with the same reference list's length (i.e. the same out degree k_{out}), and computed the average number of papers with $awt > 7$ and $dos > 0.7$ in this lists. Then I've done the same with zero SBc papers: for every out degree value I've searched for 200 papers, calculated the number of top *delayed recognition* papers ($awt > 7$ and $dos > 0.7$) in their k_{out} and computed the average. For the SB line, statistical significance is good for $k_{out} < 10$ (average done on more than 100 papers), but for $k_{out} > 15$ the average is computed on less than 30 papers. However, we can still conclude that SB papers tend to have other SBs in their reference lists with a probability higher than normal (zero SBc) papers.

.

The difference between the two distributions is even more clear if we consider only papers published before 1980: in those years, many of the actual SBs had not yet been discovered, so the papers that cite them were, generally, much less, and the fraction of these that are SB is higher (fig 33):

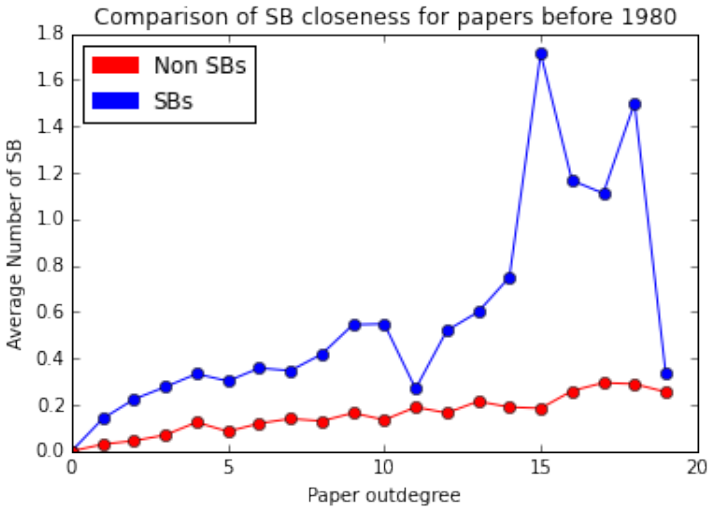

Figure 33: The difference between the two distributions is even more clear if we consider only papers published before 1980: the reason of this is that many SBs had not yet been discovered, and so for normal papers was even more unlikely to randomly cite a SB

.

7 SBs in the models of citation dynamics

In section 3 I have briefly described some of the most important models of citation dynamics.
Here I'm going to analyze them in the context of SB behavior.
As already pointed out in fig 16, both preferential attachment and the null random model are not able to describe B's distribution (and with it any other feature of SBs).
Obviously, the same can easily be said for Simkin's branching process model: since direct citations appear only to preceding year papers, the only SBs that can arise are the ones that do not have years without citations, a very small subset compared to the whole set of SBs.

7.1 SBs in the mean field approach

In section 3.4.2 we have considered the mean field approach of the redirection/copying models.
In this method, while assembling a reference list an author chooses some papers, reads and cites them, and copies from them.
The number of copied references depends on the publication year of the papers, so a huge role is played by time distribution of citations.
Now that we have an algorithm to extensively search for SBs, we can analyze them as a group: for example, speaking of time, it is obvious that SBs have citation lives that are very different from the ones of normal papers.
In the graph of Fig 34 I've compared the two time distributions: to draw it, I've taken all normal and delayed papers and calculated how many years after publication each of them gets at 10 citations.
The two distributions are very different.

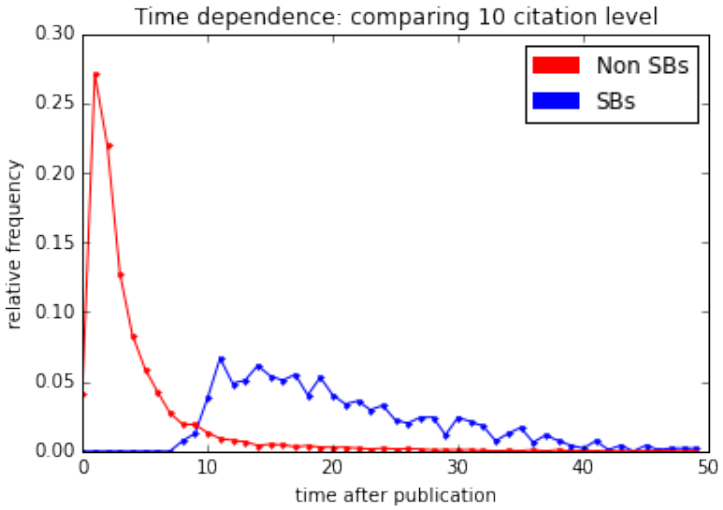

Figure 34: Comparing between time distribution for SBs (blue line) and non SB (red line) papers. Obviously, SBs, on average, take much more time to get to the 10 citation level

.

Apart from this trivial observation, we could be also interested in understanding if the life of SBs after the awakening somehow resembles the one of normal papers.
In the graph of Fig 35 for every SB on the x axis there is how many years after the awakening it took to get 10 more citations (10 cit from the level it had when it was awakened).
What is found in Fig 35 is that the two distributions are still very different, so much that the blue one is not lognormal anymore (fig36): SB papers, even after the awakening, need more time in order to be fully appreciated.

This observation leads to the conclusion that the kernel found in section 3.4.2, that described time aging of citations in the context of redirection/copying models, does not work for SB papers, even changing their year of publication with the year of the awakening.

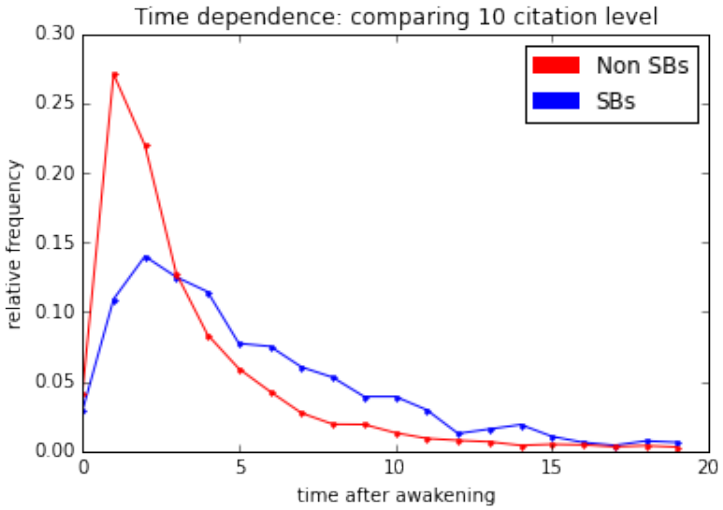

Figure 35: Comparison between time distribution of SBs (blue line) and non SBs (red line). SBs, after awakening, still need more time in order to get to the 10 citation level

.

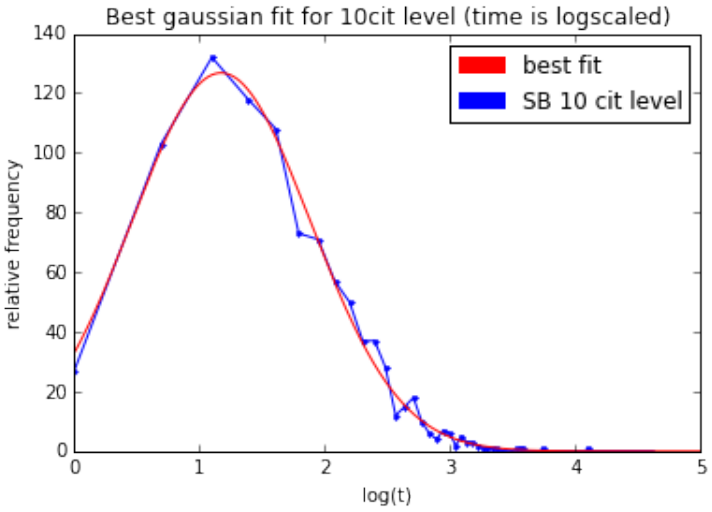

Figure 36: If the time distribution seen above follows a lognormal, when we put a log scale on the x axis we should have a Gaussian distribution. To test if the distribution seen above is still lognormal, I have scaled time and fitted data with a Gaussian function. The best fit is drawn in red: the pvalue for this fit is very low, $\sim 3\ 10^{-19}$.

7.2 SBs with Wang Song Barabasi formula

As seen in section 3.5, in the paper in ref [14] Wang et alii found a general formula to describe the cumulative distribution of citations of a paper by means of three different parameters: μ, σ and the fitness λ.
In the first place, it would seem reasonable to describe SBs by means of these three values, but this is a much harder aim than it seems.
To take a look at how this formula (eq 11)behaves in general, I've selected a small subset of 10000 papers and tried to fit them with it.
The 3D distribution of parameters is reported in Figures 37, 38,39 .

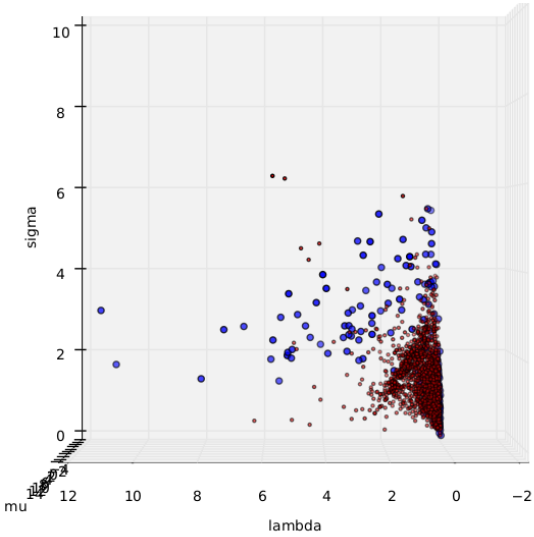

Figure 37: Relative distribution of the values of σ and λ in the subset of 10 000 papers. Red dots come from less than 200 iterations, blue ones from a number of iterations between 200 and 500

.

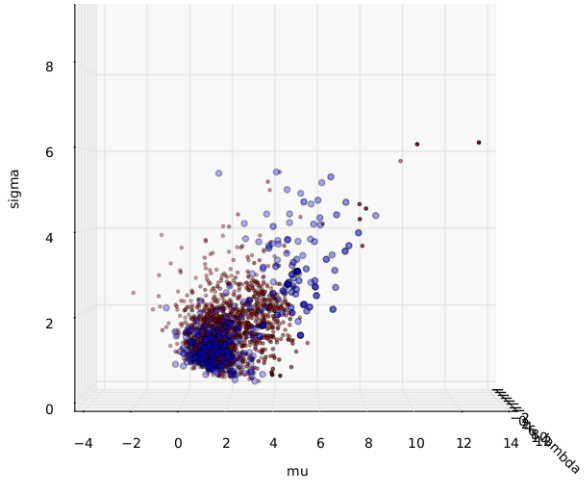

Figure 38: Relative distribution of the values of σ and μ in the subset of 10 000 papers. Red dots come from less than 200 iterations, blue ones from a number of iterations between 200 and 500

.

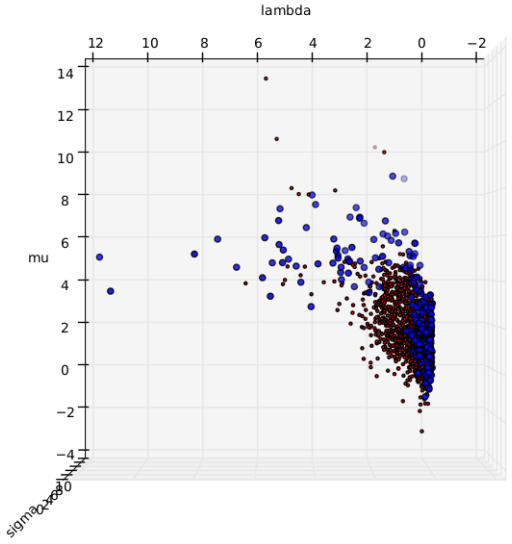

Figure 39: Relative distribution of the values of λ and μ in the subset of 10 000 papers. Red dots come from less than 200 iterations, blue ones from a number of iterations between 200 and 500

.

What emerges from this analysis is that there are a lot of papers in the SB region (high μ and relatively low σ values) for which the fit takes a lot of time to converge (the method is Nelder Mead because it gave better results. The required accuracy for parameters is 0.01).
This way, an algorithm based on WSB formula would take a lot of time to get to an accurate result, and this is not what we are looking for.
The reason why the algorithm is so slow in finding the best fit is probably due to the fact that WSB formula does not describe properly the citation distribution of SBs: if we consider the set of top SBs ($awt > 7$ and $dos > 0.7$) that have gotten more than 25 total citations, we have that (see fig 40 for an example) a 90% of them has a p-value under 0.6, with a 40% under 0.1 !

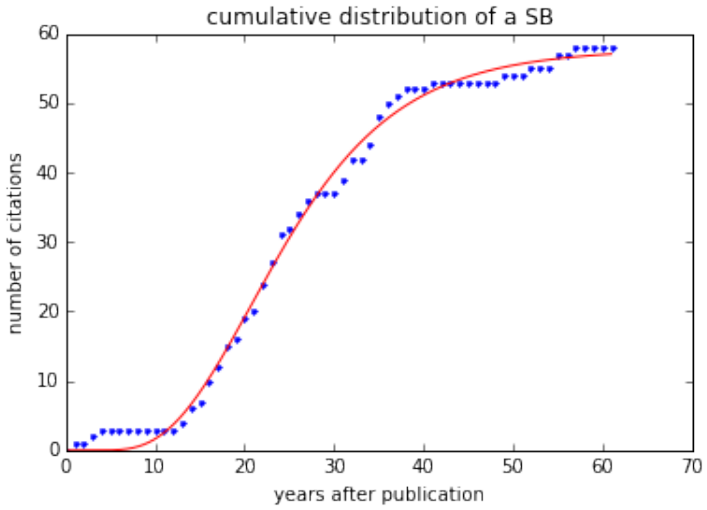

Figure 40: Cumulative distribution of a top SB paper. The red line is the best fit with the Neldel-Mead method. The p-value for this fit is very low, $\sim 5\ 10^{-8}$

.

8 Studying the 'Prince Hypothesis'

Together with the definition of SBs, in its paper [16] Van Raan introduces also the interesting idea of the prince, i.e of a paper that has the ability to induce citations on another.
The definition of SB given there, together with the defect of being too strict, had also the merit of finding the prince in a very straightforward way: it was the first paper to cite the SB after the stop.
In our case this naive definition is not self consistent : a lot of papers with non zero SBc do not experience any period of stop, so that by the time of the change in the rate of citations we can not univocally determine who's the paper responsible for it.
A good idea to find this *prince* is to use some of its expected topological properties based on the mechanisms of the models in section 3.4.
As we had seen there, almost all of them focus on constraining time distribution and triangles: so to have big amounts of citations one needs a relatively new paper.
The details of this hypothesis are discussed below.

8.1 Are SBs awakened by super cited papers?

As already pointed out in the paper [18], preferential attachment does not explain the existence of SBs: citation to older papers can happen only if a paper is already very famous.
In the model by Simkin (Ref [12]) papers may be added in reference lists by copying from lists of other papers: this could be the phenomenon responsible for creating triangles in the network, and even though the model can not generate SBs (mainly because direct citing is allowed only for publication of the year before), this idea can be easily tested.
Let's suppose the majority of references are added simply by randomly copying other citations. If a very good paper (one that has gained a lot of citations in its lifetime) cites a paper published many years before, there is a very high probability that a lot of new papers copying from the first paper will also cite the second. This lucky paper would get all of a sudden a lot of citations, becoming a SB.
In this scenario, the reason why SBs arise would be quite futile: their delayed life is just a matter of chance, not the demonstration of any intrinsic scientific value.
To test this hypotheses I've selected all top rated SBs ($awt > 7$ and $dos > 0.7$, they are almost 1500) and considered all the citations they gained in a time window of 7 years around the awakening time (from $awt - 2$ to $awt + 5$).

Then I analyzed the in-degree of all the citing papers in this time interval and took the highest one. Then I did the same for papers with $SBc = 0$: I took a time window of five years after publication and computed the highest in-degree of citing papers.
Values on the x axis are the total in-degree of papers, while on y there is the average highest in-degree of papers in the time windows described above (Fig 41)

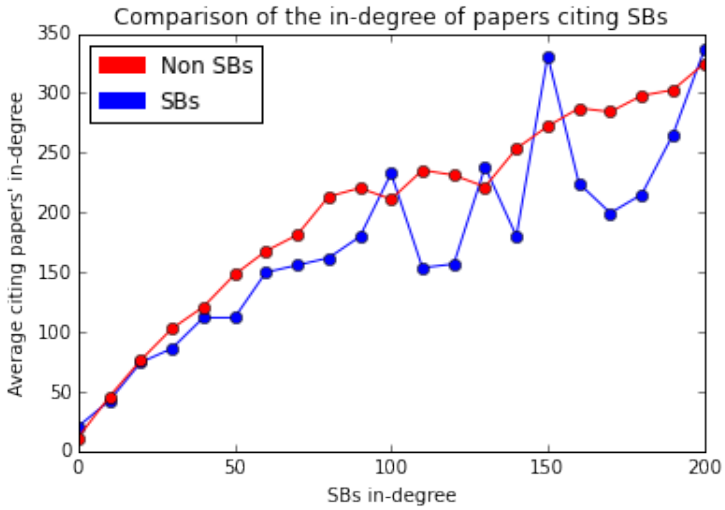

Figure 41: Comparing distributions of in degrees of papers citing SBs. I have grouped SBs in-degrees in bins, so that statistical significance is good for the first eight (they are averaged on more than 80 SBs each), but quite low for the others.

.

From this graph, it is clear that SBs are not awakened by super cited papers.
However, we must not forget that we are considering only APS papers: if such a prince was published in another journal (a journal with a higher impact factor, as pointed out by Braun in [17], we would only see the strange situation in which a lot of low cited papers independently cite the SB many years after its publication.
Again, this does not seem to be the case: until 1980 papers published in the APS where very technical and specific of physics (quarks, radiation, etc), so it seems reasonable to think that interdisciplinarity at the time played a minor role.
I analyzed the behavior of SBs awakened before 1980, and, again, it is not explainable just with the random citing hypothesis (since we have much less SBs, statistical significance is lower but still enough to say that there is no super cited paper) .

8.2 Do SB and prince 'marry' after kiss?

When writing a new scientific article an author borrows ideas from previous works and uses them in order to come up with something new.
When reading this paper, another researcher can discern articles that are put as simple references from the ones that are fundamental. So, even if not reading the entire articles in reference lists, some of them are more likely to be cited than others.
Another possible explanation for SB awakening is that the prince is somehow a paper that explains to scientific community the ideas already developed in the SB (evidently, SBs are too hard to be understood at the time of publication).
If so, the two papers must have a very similar citation history, but, again, this is wrong: as pointed out by Brawn in 2010, prince and SB have completely different citation lives after the 'kiss'.
But, if not for a long time, one still expects their lives to be very similar, at least immediately after the awakening. To verify this hypothesis, I've used a coefficient, that I called P_c1 (Sorensen-Dice coefficient in the bibliography).
It can be computed for any pair of papers and quantifies how closely related are their lives in the 5 yeas after publication of the latest.
More precisely:

$$P_c1(a \to b) = \frac{2\ N_{a,b}}{N_a + N_b}$$

where $a \to b$ means that a is published after b, $N_{a,b}$ is the number of papers that cite both a and b in the 5 years after publication of a, N_a is the number of papers citing a and N_b is the one for b in the time interval between (publication year of a) and ((publication year of a)+5).
P_c1 varies in the interval $[0, 1]$. In order to be statistically significant, I took only P_c1 values for papers that have $N_a > 5$: an average taken on just 1 or 2 papers would not give much information.
For each of the top class SBs ($awt > 7$ and $dos > 0.7$) and for every paper that cites them in the interval of time $awt - 2 < t < awt + 5$ I've calculated this coefficient.
Afterwards, to every SBs I've associated the maximum P_c1 value from the array. Then I've done the same for non SB papers, calculating the maximum P_c1 value in the 5 years after publication. The graph of fig 42 compares the two distributions.

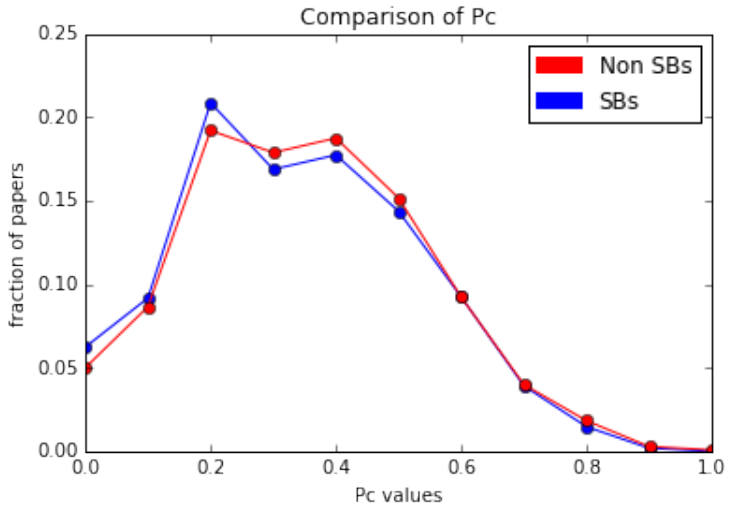

Figure 42: Comparison of the maximum of P_c1. The red line is computed on more than 75000 papers, while the blue one on SBs with $awt > 7$ and $dos > 0.7$. The two distribution do not show any significant difference.

.

As is clear from it, the two distribution do not show any significant difference, and this can not be explained with the idea of a prince published in other journals: in fact, if we compare the distribution of non SB papers with the one of the SBs awakened before 1980 (period in which is reasonable to think that interdisciplinarity did not play a significant role), the 2 distributions are still very similar (fig 43).

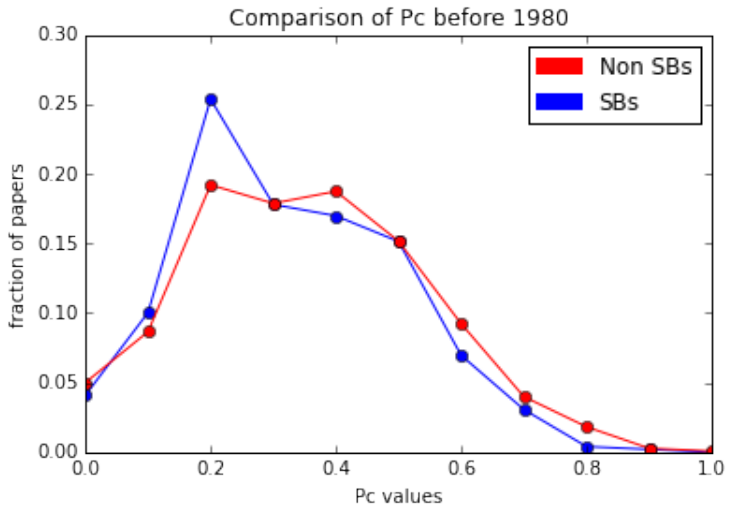

Figure 43: Comparison of distributions of the maximum of P_c1 with SBs awakened before 1980 (with $awt > 7$ and $dos > 0.7$). Again, the two distributions appear to be very similar.

.

I also tried to analyze the dependence of the P_c1 coefficient from the in-degree of the SB (Fig 44):

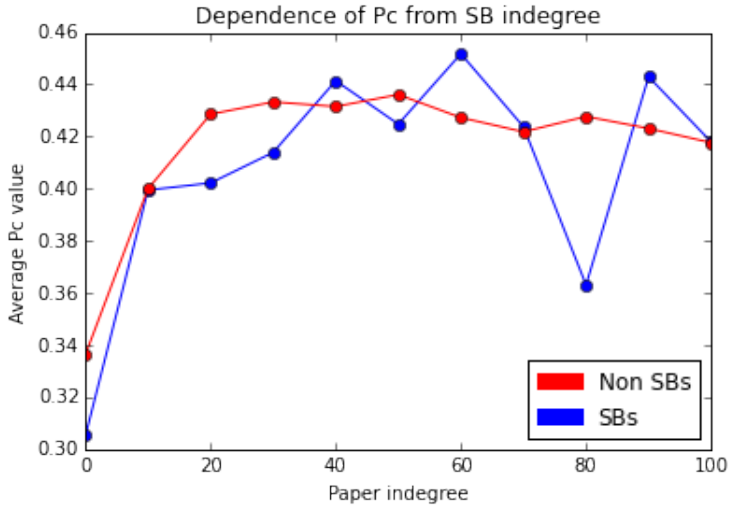

Figure 44: Dependence of the average maximum P_c1 value for SB and non-SB papers. I grouped papers in bins, so that averages are calculated over more than 100 papers for the first 3 bins, over less than 30 for the last 3 (so these ones may not be very significant). Anyway, P_c1 values for SBs are very close to the non-SBs ones.

.

The only significant change in the two distribution is that for SB papers with less than 30 citations the maximal P_c1 value is on average less than normal papers' ones, exactly the opposite one would expect if SBs' lives are heavily dependent on one article.

I tried changing a little the definition of the coefficient, to see if something changes (this coefficient is called cosine similarity in the bibliography):

$$P_c2(a \to b) = \frac{N_{a,b}^2}{N_a \ N_b}$$

and, again, I calculated it for SBs and non SBs in the 5 yeas after publication. The graph of Fig 45 shows, again, no substantial difference.

The analysis above seems to deny that the actual scientific content of

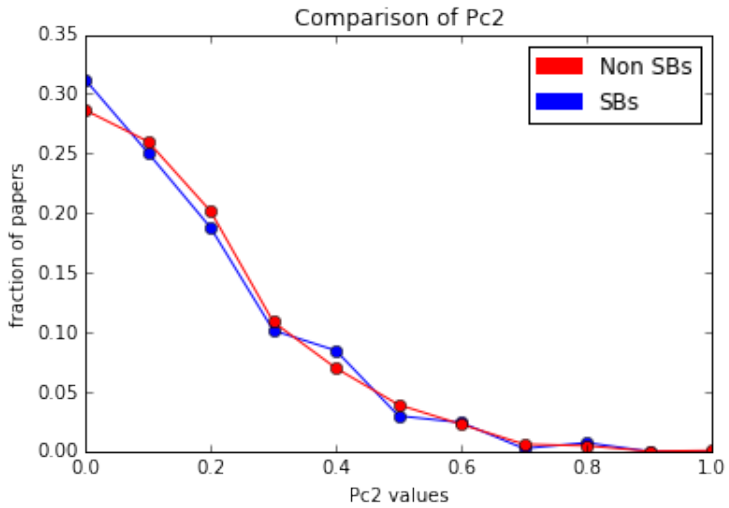

Figure 45: Comparison of distributions of the maximum of P_c2 . Again, no substantial difference.

.

princes' papers is a simple rewriting of SB's ones: every paper citing the SB actually adds something to scientific knowledge so that it is not automatic to cite prince and SB together.

8.3 Are SBs citations coming from just one article?

In the previous section I've analyzed the idea that prince and SB are very closely related papers, and it does not seem to be correct.
A SB is a paper that, somehow, is forgotten for many years, and that at a certain point gets a lot of attention from scientific community. It would be very surprising and strange to find that all these citations come independently, without citing each other: there must be a common hub in the network.
In order to understand a little bit more of what happens I've considered a third coefficient:

$$P_c3(a \to b) = \frac{N_{a,b}}{N_b}$$

It can be calculated for any couple of papers a and b (b is published before a). $N_{a,b}$ is the number of papers citing both a and b in the 5 years after the publication of a, while N_b is the total number of papers citing B in the same period.
I took P_c3 values for all the papers in the interval $awt - 3 < t < awt + 5$ for all the SBs with $awt > 7$ and $dos > 0.7$, calculated the maximum value in the array of P_c3 and compared this distribution with the one of non SBs (Fig 46).

Surprisingly the two distribution are quite different, especially around high values. This difference appears much more clear if we separate SBs discovered before 1985 (Fig 47) and after 1995 (Fig 48)(i.e. if the sum of the publication year and awt is less than 1985 or more than 1995):

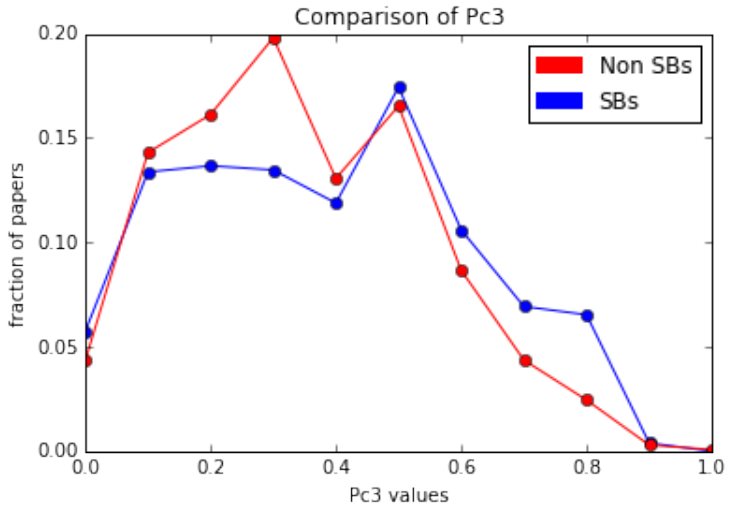

Figure 46: Comparison of distributions of the maximum of P_c3 .There is a difference in the values of the distributions, especially around the peak and for $Pc3 > 0.7$.

.

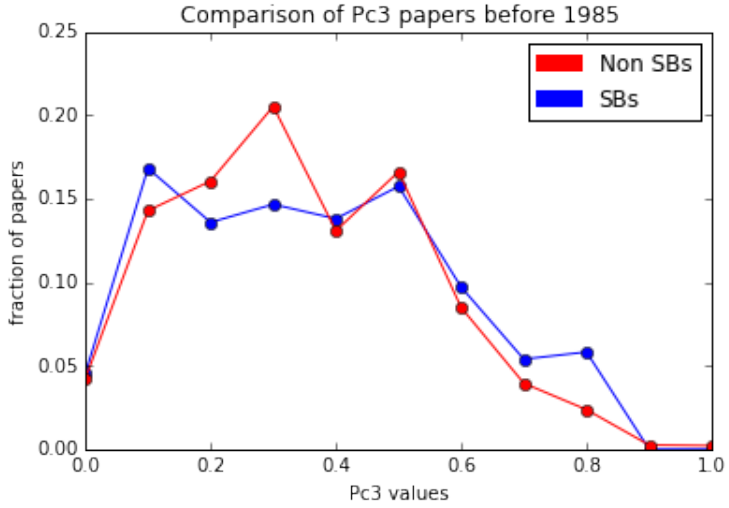

Figure 47: Comparison of distributions of the maximum of P_c3 .The distribution in blue comes from the analysis of more than 300 SB papers discovered before 1985

.

From these two graphs we can conclude that in the years immediately after their awakening SBs very often depend from the citations of just one

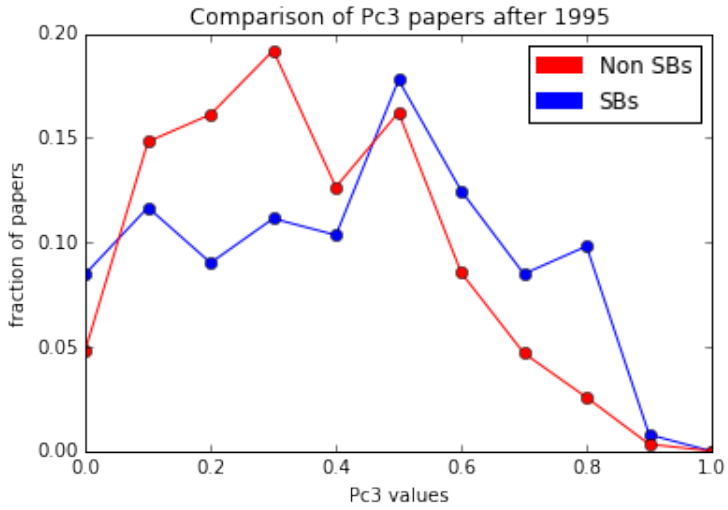

Figure 48: Comparison of distributions of the maximum of P_c3 .The distribution is calculated on more than 400 SB papers awakened after 1995

.

paper, that we could call the prince.
It has to be noticed, however, that a very large fraction of SBs do not have such a paper.
Moreover, there is quite a big difference between the distributions before 1985 and after 1995: having to put much more papers in reference lists (remember that the length of ref lists grow exponentially with a parameter $\sim 0.02yr^{-1}$ as seen in section 3) citations given after 1995 create much more triangles.
On the other hand, in the same time interval the number of zero P_c3 values has increased: this may be the effect of interdisciplinarity: for these papers there is another paper,published in another journal, that is acting as a catalyst of citations for the SB, and we can not see it.

Finally, the dependence of Pc3 from the in degree of SBs (Fig 49):

The graph shows how the SBs that are more dependent on just one paper are the ones that have gotten less citations in their life

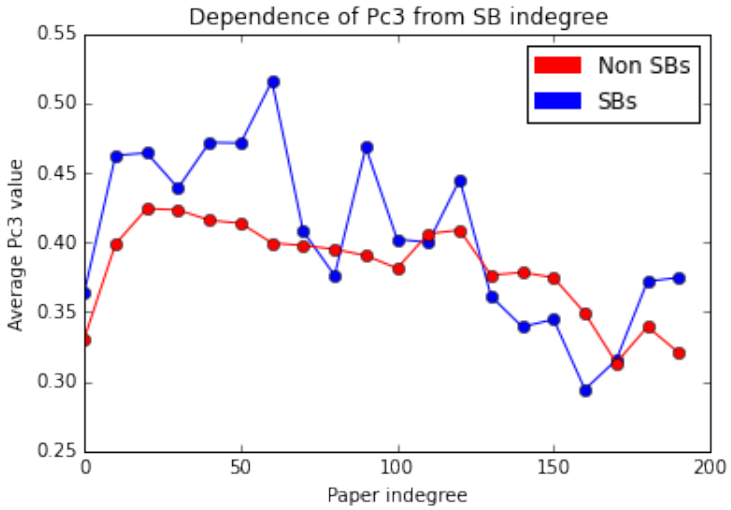

Figure 49: Dependence of the average maximum P_c3 value for SB and non-SB papers. I grouped papers in bins according to their in degree. Again, averages are calculated over more than 100 papers for the first 4 bins, while for the others we have less than 30 from in degree 80 onwards. For the first bins average P_c3 values are quite different from the ones of non SB papers

.

9 SB grouping

In sec 6.1 we saw that SBs tend to be close to each other.
This can be reasonably explained if we consider that at the awakening of a paper new attention is given to all the papers somehow related to it (that share authors or keywords).
In order to study this phenomenon I've grouped SBs considering related the ones that are neighbors and that were awakened more or less at the same time. More precisely, for each SB (SB_1) with $dos > 0.7$ and $awt > 7$ I've searched for other SBs (SB_2) in their reference lists with $Ya_1 - 3 < Ya_2 < Ya1 + 3$, where Ya_1 is the awakening year of SB_1 and Ya_2 the awakening year of SB_2. The clusters that are formed this way have papers with very similar citation histories, and we can examine some of their basic properties, such as time distributions. Fig 50 and Fig 51 show the publication year and discovery year (i.e. publication year plus awt).
I grouped SBs into three groups based on the number of papers within the cluster.

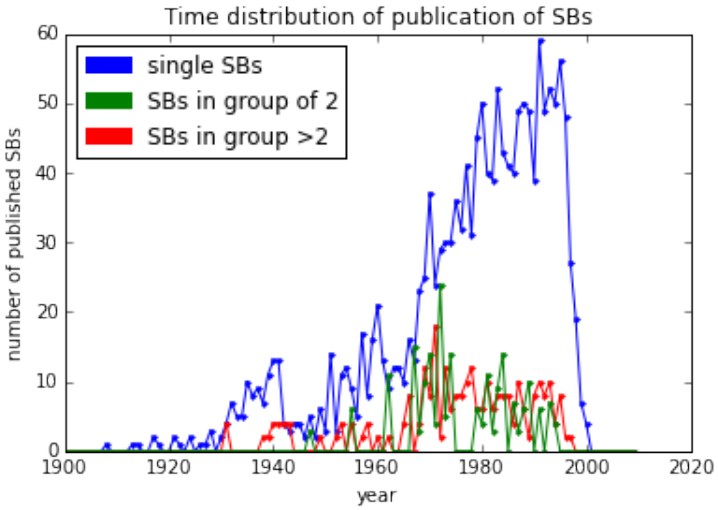

Figure 50: Blue line describes SBs not in clusters, while the green and red ones describe couple and groups of more than 2 SBs, respectively. Single SBs are clearly the majority, and tend to grow in number with time, while the other two from 1940 are more or less stationary. Note that around 1940 there is a peak: during the second world war it was forbidden for scientists to give information on their research, so after it scientific community had to put a strong effort in trying to rebuild a dialogue

.

As can be noticed from Fig 51, quite interestingly clusters of SB, even though published regularly from 1940 onwards, were discovered only after 1980s.
One reason for this may be that with physics becoming interdisciplinary many papers were revalued, or maybe the developing of instruments and technical precision helped entire topics to gain new attention.
Clusters of SBs are very interesting also because they manifest the appearance of a completely new branch of research. Examining them, we are interested in answering some basic questions, such as: if some SBs are awakened in the same year, do they share the same prince?
To answer this question, and in order to be more precise, I'm going to analyze some examples.

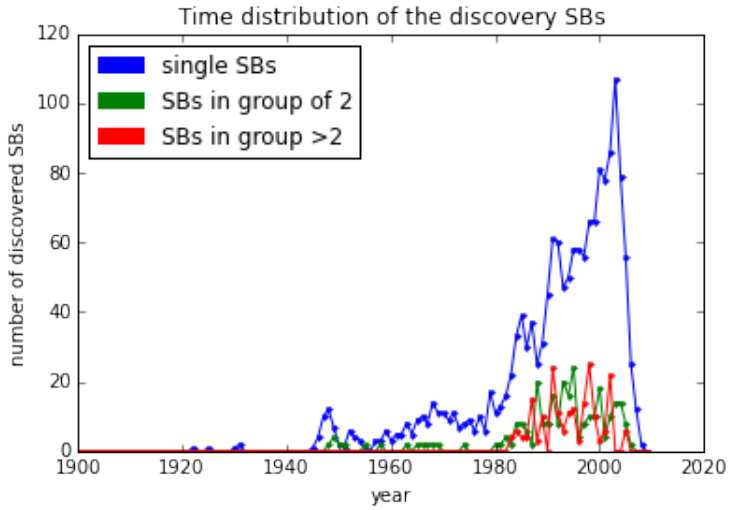

Figure 51: The number of SBs discovered every year tends to grow with time, starting from more or less 1940, while clustered SBs tend to be awakened only after 1980s, even if many of them had been published much time before. Note also the peak in the number of SBs discovered immediately after the war, when the flow of information started again to be free

.

9.1 SBs group dynamics: examples

9.1.1 Spontaneous breaking of Lorentz symmetry in the Standard Model

The first example comes from pure theoretical physics: the spontaneous breaking of Lorentz symmetry in the Standard Model (Fig: 52): the cluster is formed by 4 SBs, published at the end of the 80s, and thanks to the relative scarcity of papers, the analysis is quite easy to accomplish.

In 1986 the paper 'Spontaneous breaking of Lorentz symmetry in SM' ('PhysRevD.39.683') presented what was an inconsistency within the formalism of the standard model. Despite an effort in trying to solve this issue, manifested by the publication of other 3 related papers (the other SBs) immediately after the first one, no convincing advancement was made until 1998.

In the work by Colladay 'Lorentz Violating extension of standard model', of 1998 (PhysRevD58.116002), new mathematical tools were introduced in order to extend the SM: this clear and convincing work brought a large number of new citations, and the whole set of SBs became very popular. As can be seen from Fig 53 the 4 SBs were actually awakened by the same paper.

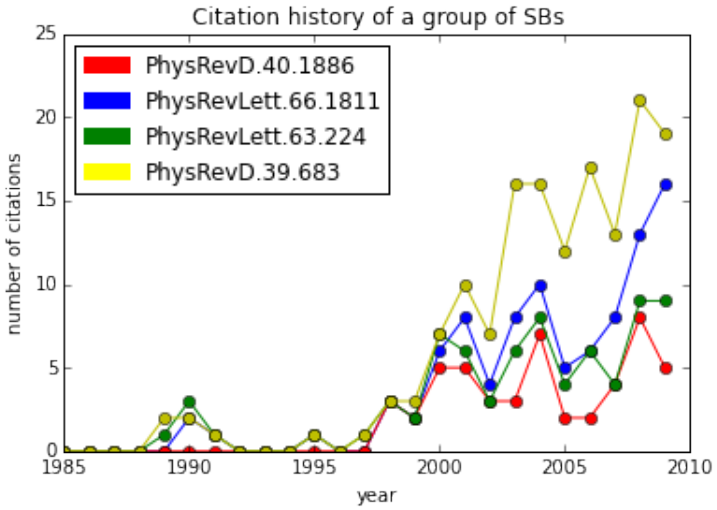

Figure 52: Citation History of a the group of SB related to the breaking of Lorentz symmetry. Published at the end of the 80s, these 4 papers did not get almost any attention since 1997

.

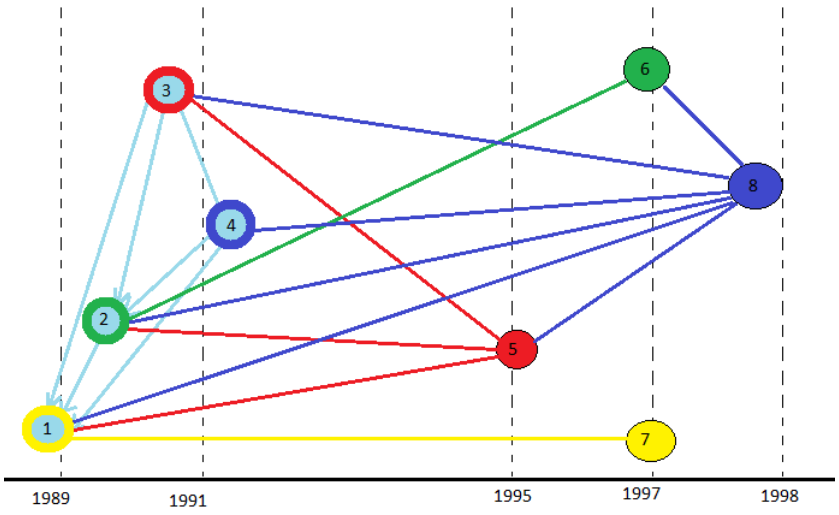

Figure 53: Representation of the citation history of this group of papers. The four papers in light blue represent the SBs, and the circles have the color used in fig 52 to represent their citation lives. Little attention was given to them until the publication of paper 8, in 1998

.

However, if we use the naive definition of prince as 'the first paper to cite the SBs after the break', we would end up choosing one of the papers published in 1997, while the paper responsible for the awakening was published in 1998.

9.1.2 Exchange Anisotropy

In 1956 a paper called 'New Magnetic Anisotropy' by Meyklejohn and Bean was published, with the aim of describing some results on the analysis of the interaction between antiferromagnetic and ferromagnetic materials.
What was found was a new type of anisotropy that they called 'Exchange anysotropy'. In 1957 a very similar paper (same authors, same title) was published that introduced a new simple model for this phenomenon.
Despite an initial attention, the two works were almost completely ignored since the late 80s, when other 2 papers were published, pointing out that the old model was unable to completely describe the system.
From my data, that include only APS journals, after these two there is another period of break, but this is incorrect: in 1997 the paper 'Calculations of Exchange Bias in Thin Films with Ferromagnetic/Antiferromagnetic Interfaces' by Koon reports a brief history of the topic.
It is stated that ' in recent years exchange bias in thin films has found important technological application in such devices as magnetoresistive sensors', and since the fundamental origin of the phenomenon was still unclear, a lot of research had started since 1987, and had continued for all 1990s: analyzing tha reference list of this paper I found other works published between 1993 and 1995, in other journals.
In the end, the interest in the topic had started after the usage of films in technological applications. No published paper was actually responsible of the awakening of the SBs, but an external phenomenon, not encoded in citation dynamics.
Another observation I want to mention is that some of the links from the paper by Koon to other papers of the network were missing from the APS data set!
This archive errors, together with the intrinsic difficulty of the phenomenon, renders a more systematic analysis almost impossible (Fig 54, Fig 55).

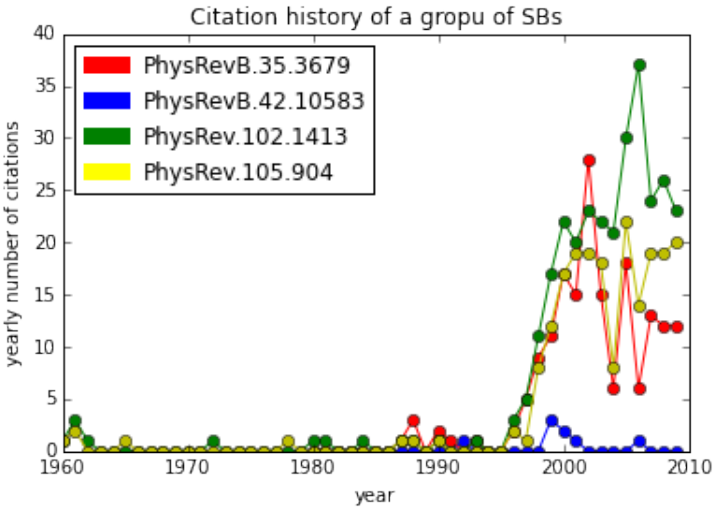

Figure 54: In APS dataset all four are SBs, while the actual number is only two since the ones published in 1987 and 1990 had not been sleeping at all: citations had come from other journals

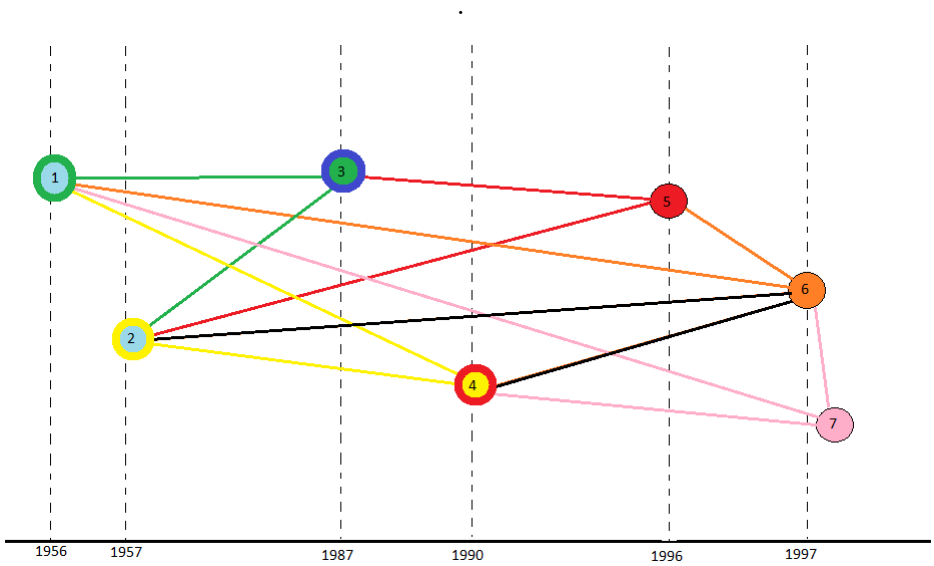

Figure 55: Representation of the history of the papers. Again, the ones in light blue are the SBs, and the colors of the borders are the same used in fig 54. No one of the papers published after 1987 is alone responsible of the awakening of the SBs, and the black lines are citations reported in the ref list but mysteriously missing from APS dataset.

9.1.3 Ferromagnetic Compounds of Manganese with Perovskite Structure

In 1951 a paper called 'Interaction between the d-Shells in the Transition Metals. II. Ferromagnetic Compounds of Manganese with Perovskite Structure' was published. In that period an empirical relationship between electrical conduction and ferromagnetism in Perovskite lattices of manganese had been discovered, and in this paper the authors try to explain the phenomenon using a 'double exchange process'.
Papers published the years after are results of experiments that confirm formulas and hypothesis of that first paper. Very little attention was given to the topic for almost 30 years, until the end of 1980s. However, it is only with the paper 'Double exchange alone does not explain resistivity of LaMnO3', that substantially says that the model proposed in 1951 was not complete, that SBs started to become very popular.
Even though not citing them directly,this last mentioned paper is without doubt the prince of all the six SBs. This is a clear sign that papers can influence lives of other papers even without citing them directly:this way models that consider only triangles and time difference to describe citation dynamics cannot take into account some fundamental characteristics of SBs (such as the tendency of being awakened together) (Fig 56 and Fig 57).

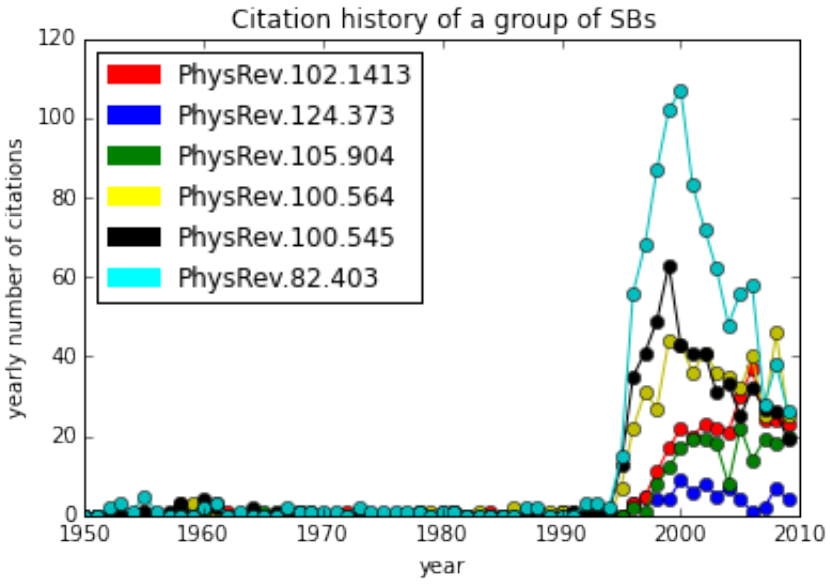

Figure 56: Another group of linked SBs. They were all published around 1950, and waken up around 1995.

.

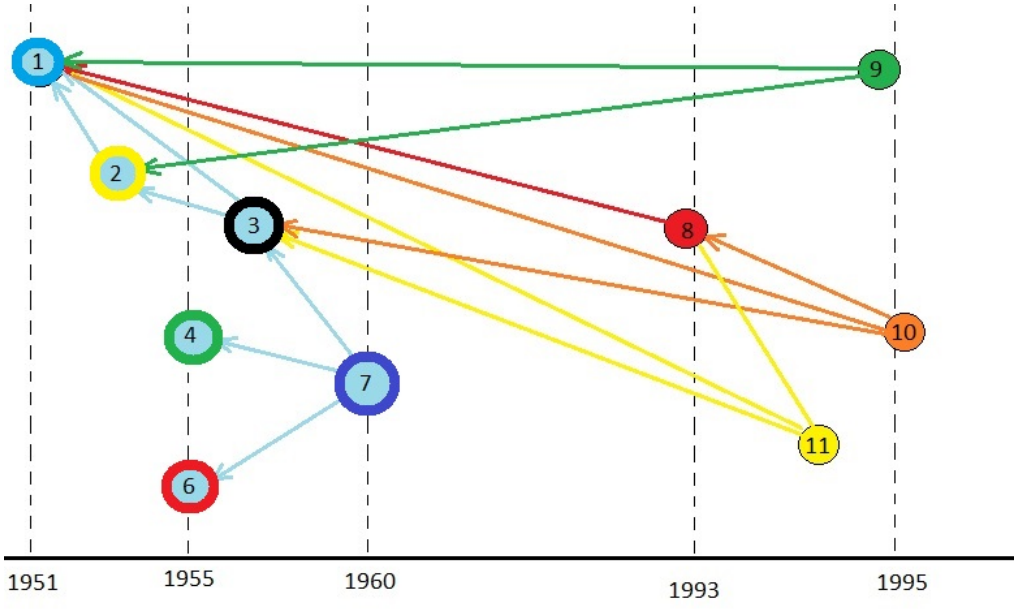

Figure 57: Representation of the history of the topic. Light blue nodes are the SB and the color of the border is the same as the color used in 56. The true prince for all these SBs is the paper represented in green, but it only cites 2 of the 6 SBs.

10 Simulation of citation dynamics with Markov Chain Monte Carlo method

In section 3 we have seen some of the most important models of citation dynamics. Redirection/copying ones focus on trying to recreate the triangles distribution, while the other models tend to put more emphasis on time distribution.
However, some of the most common and basic characteristics of SBs can not be taken into account by simple model like these.
The presence of an actual prince (responsible for the second life of a SB) would justify the idea of focusing only on triangles in citation dynamics. But in sec 8 we have seen that the great majority of papers do not have it.
In sec 9 we have also seen that a paper (that may be considered the prince of a SB) is able to influence the life of another even not citing it directly, but this mechanism is completely beyond the scenarios of the model considered till now.
In this section we would like to test the hypothesis that the total number of triangles alone is sufficient to describe all the topological features of SBs.
To do so, we will simulate the network using MCMC methods (see Appendix).

The first step to create this new fake citation network is to take only papers published before 1936, together with all the ones that cited them (even if published after 1936), and all the citations among these. The choice of picking only a subset of all the data is due to simulation time restrictions: if we wanted to use all the papers the process would take literally years.
Secondly, we have to decide what to preserve of the real network in the new one.
We will constrain:

- time difference between the extremes of each edge, i.e. the years of citing and cited articles
- for each paper, in and out degree
- the total number of triangles

Note that the first two points together do not mean we are constraining also time distribution of citations: for every paper we will keep the total number of citations but change when they are achieved.
Among all the types of MCMC methods, we will use the one called Simulated annealing: the name comes from metallurgy, where it is seen that a metal

produced with a slow cooling (annealing) is much stronger than metals produced with a fast decrease of temperature. The idea is to start looking for states on a large scale by using a high 'temperature': doing so we allow the system to freely search within the state space S (all allowed states can be sampled).
At each temperature, the system will change according to MCMC rules and to a function that has to be minimized (the energy E). Then T is slowly lowered (annealing) until we reach T=0.
The advantage of using MCMC simulations is that we get a new 'fake' state just by modifying the real one, and this is much faster than creating a brand new one.
In more detail, the algorithm:

1. starts considering a first edge in our network.

2. This edge, that goes from paper 1 to paper 2, begins in year Y_1 and ends in year Y_2. Since we want to preserve time distribution of edges, we will search for a third paper within the ones published in year Y_2 that are not cited by paper 1 (we will call this third paper 3).

3. Once we have found 3, we randomly choose within its citing papers a fourth one (4), paying attention that 4 does not already cite 2

4. The proposed move is to break the links $1 \to 2$ and $4 \to 3$ and to substitute them with $1 \to 3$ and $4 \to 2$. This way we preserve time distribution, total number of papers and of edges.

5. We will accept this move with the rules of Metropolis-Hastings algorithm, using as energy the total number of triangles. If we consider the adjacency matrix of the network (σ_{ij}), we have that the Hamiltonian will be
$$H = \frac{1}{T}\left(\sum_{i,j,k} \sigma_{ij}\sigma_{jk}\sigma_{ik} - N_{tr}(0)\right)^2$$
where $N_{tr}(0)$ is the number of triangles in the real network, and i, j and k run over all the possible nodes of the system.

6. At each passage it will not be necessary to calculate the total number of triangles: we can simply consider the difference between the one of $1 \to 2$ and $4 \to 3$ and of $1 \to 3$ and $4 \to 2$ with respect to the passage before.
More precisely, let $N_{tr}(1 \to 2, 4 \to 3)$ be the sum of triangles around

edges $1 \to 2$ and $4 \to 3$, and $N_{tr}(1 \to 3, 4 \to 2)$ the sum of triangles of the trial move. If $N_{tr}(N-1)$ is the number of triangles at time $N-1$, in order to accept or reject the trial move we will consider:

$$a = exp\left(-\frac{1}{T}\left[N_{tr}(trial\ move)^2 - N_{tr}(before)^2)\right]\right)$$

where

$$N_{tr}(trial\,move) = N_{tr}(1 \to 3, 4 \to 2) - N_{tr}(1 \to 2, 4 \to 3) + N_{tr}(N-1) - N_{tr}(0)$$

and

$$N_{tr}(before) = N_{tr}(N-1) - N_{tr}(0)$$

and if the number of triangles of the trial move is closer to $N_{tr}(0)$ than $N_{tr}(N-1)$ the move is accepted, otherwise one considers a random number (as in Metropolis Hastings method), and if this number is higher than a, the move is rejected, and the system stays in the configuration of time $N-1$ also at time N.

7. We will firstly consider very high temperatures ($T \to \infty$) so that we will accept all possible moves. In the end, we will have a completely uncorrelated configuration with respect to the initial one. Then we will slowly decrease T as N_{tr} gets closer to $N_{tr}(0)$

The problem of this algorithm as is that it is very slow (it takes really a lot to get to a reasonable configuration).
We don't have to forget that, as already pointed out in section 3.3.1, the network of citations has a lot of triangles and papers related to common topics tend to cite each other in the 99% of the cases. So the difference in the number of triangles for the random case and the real one is very big ($\sim$ 80000 triangles).
In order to make the algorithm a little bit faster, at point 2, instead of considering as possible paper 3 all the papers published in Y_2, we will consider a more sophisticated mechanism: of all the papers in the reference list of paper 1, we will take all the papers that cite and are cited by them and put them in an array of possible choices. Then we take all these papers we have written in the array and, again, add (to the same array) all the papers that cite and are cited by them.
This procedure does not violate MCMC requests: the matrix of the trial moves Φ can be arbitrary as long as the Markov chain is regular and satisfies detailed balance.

10.1 MCMC simulation: results

I have run the above algorithm until I found more or less the same total number of triangles as the real network (the difference between the two is ~ 700 triangles, while for the random case the real one had nearly ~ 80000 triangles more).It took me a total of almost 300 hours
Then I have compared the results for the two systems, recalculating many of the distributions already shown in the other sections.
The first comparison is between the distribution of triangles around edges: every edge in the network is normally part of 1 or more triangles, and for every link I've calculated it in the real and simulated case (fig 58).

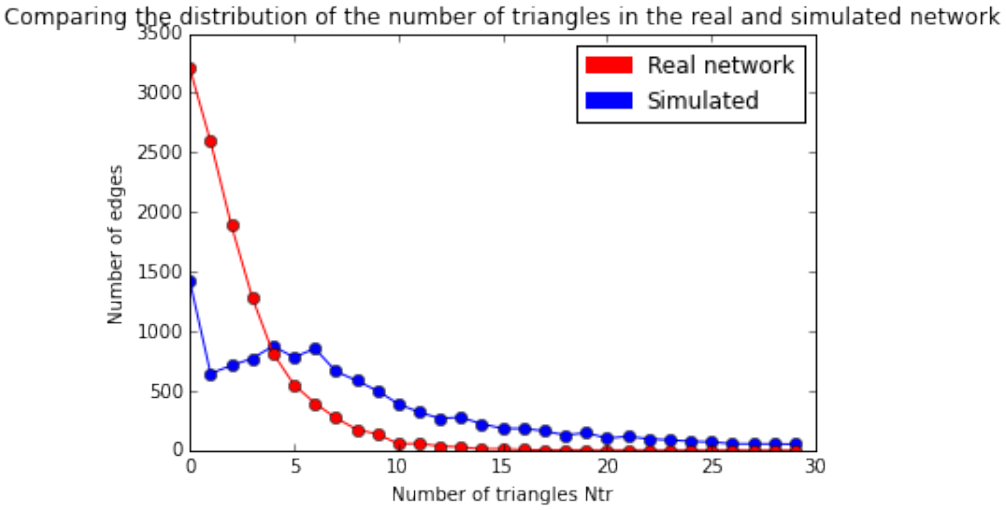

Figure 58: Distribution of number of triangles around edges of the networks. Red dots stand for the real distribution, blue ones for the simulated. On the x axis are the number of triangles, on y the number of edges that have that specific $N_{tr}(a \to b)$ around them

.

As can be easily seen, the two distributions are very different, and this is mostly due to the choice of the matrix of trial moves: in our system we have considered all the states with the same number of total triangles (N_{tr}) equally probable, but obviously this does not mean we are constraining also the other quantities of the network.
What we have in the simulated case is a clustered network in which papers tend to be very close to each other, much closer than the real case.
Said so, it is obvious that also all the other simulated distributions are very different from real ones (fig 59 and 60):

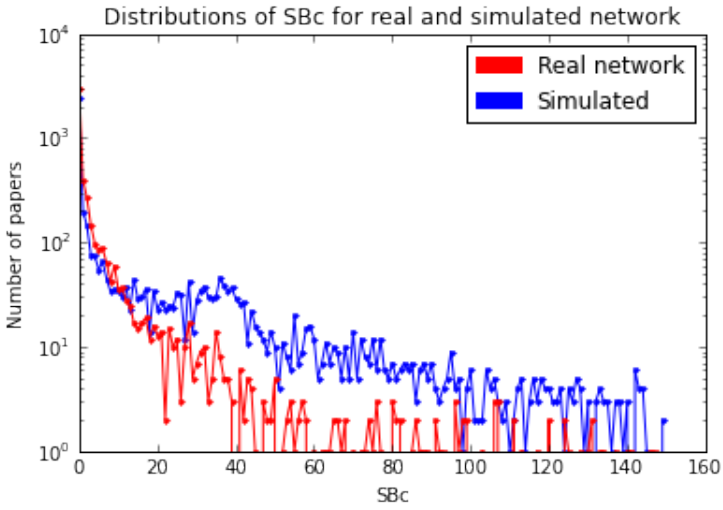

Figure 59: Distribution of SBc in the Simulated and real case. The two lines are very separated: in the simulated case, on average, we have much higher values of SBc. The scale on the y axis is logarithmic

.

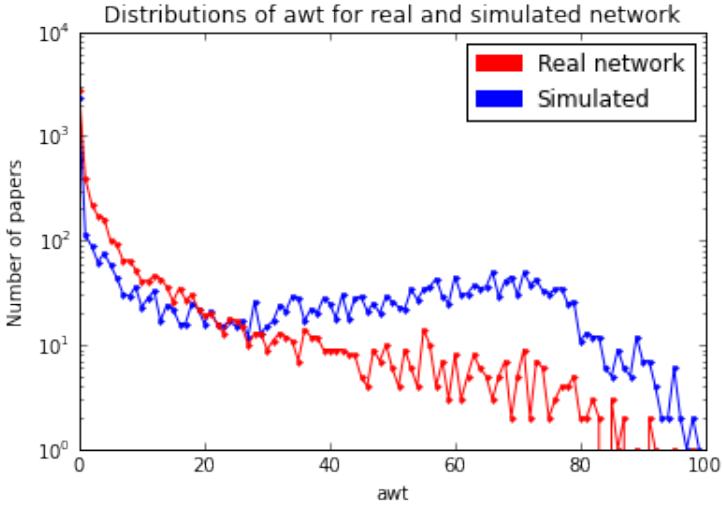

Figure 60: Distribution of awt in the Simulated and real case. Again, the two lines are very separated: even for the awakening time, our null model preserving time and triangles gives on average much higher values than real one. The scale on the y axis is again logarithmic

.

Even the number of top class SBs ($awt > 7$ and $dos > 0.7$) increases in the simulated network, passing from 453 of the real case to 971 (remember that we are considering only papers published before 1936).
The distribution of the publication year of top class SBs is reported in fig 61:

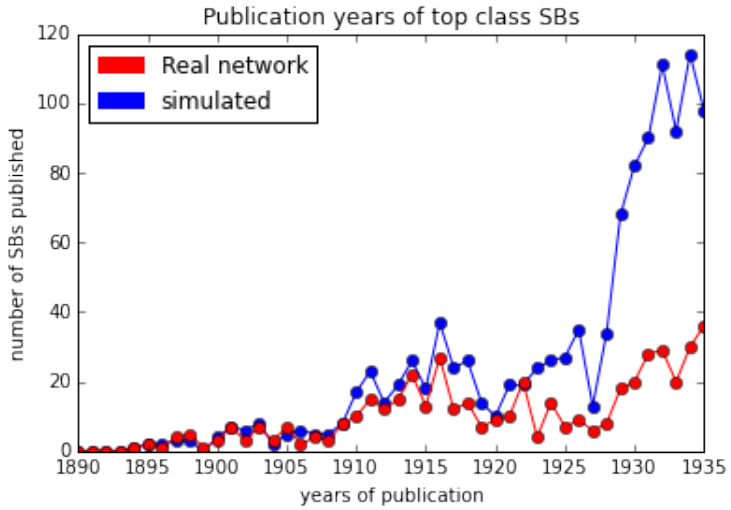

Figure 61: Comparison between the publication years of top class SBs in the real and simulated network. As already mentioned, in the simulated case, the total number of SBs is much higher, so the areas under the lines are not equal.

.

Another distribution of interest may be the dependence of triangles from time difference of the starting and ending nodes of edges: the graph in fig 62 is created considering all the edges in the system, grouping them for the time difference between their extremes and, for each of these bins, calculating the average number of triangles N_{tr} around each.

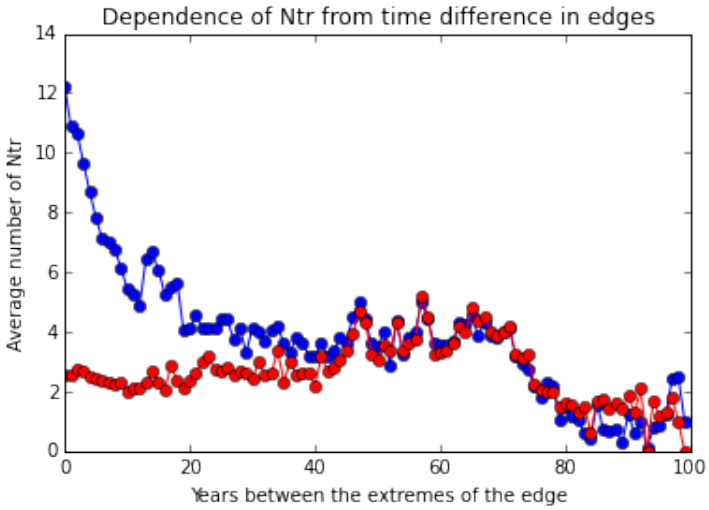

Figure 62: Comparison of the number of triangles N_{tr} around edges and time difference between their extremes. As can be seen, especially for small x values, N_{tr} in the simulated case is on average much higher than the real one.

.

However, if we consider the closeness of SBs and their tendency of being clustered, we have (Fig 63) that there is no clear difference between normal and SB paper: in our simulation SB do not tend to be close to each other, while in the real case, as can be seen in fig 64, the difference is big.
This is due to the fact that we have focused only on preserving triangles, while in the awakening of SBs the attention goes beyond its neighbors and it can not be described only by simple redirection/copying methods.

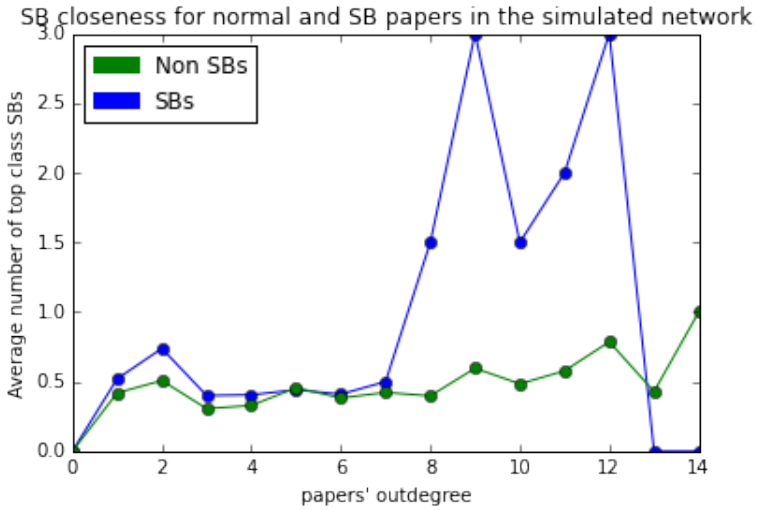

Figure 63: Comparison between closeness with top SB papers ($awt > 7$ and $dos > 0.7$) for SB (blue dots) and normal papers (green dots) in the simulated network. Considering that for out degree more than 8 in the blue line the average is computed only on 1 or 2 data, we can conclude that there is no significant difference between the two distributions

.

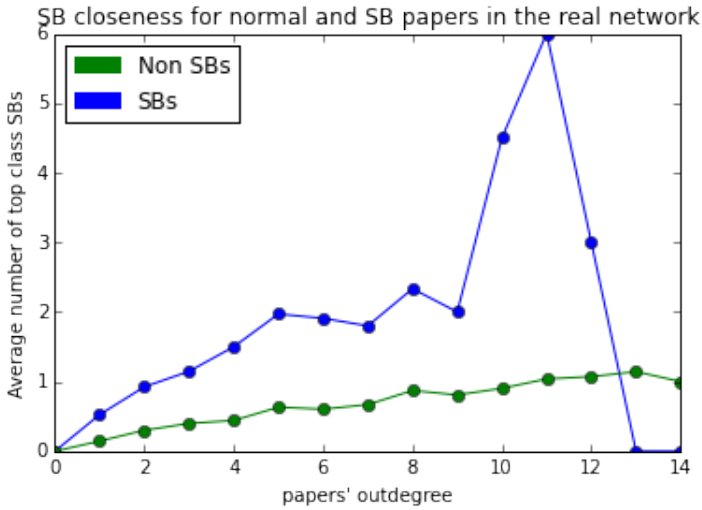

Figure 64: Comparison between closeness with top SB papers ($awt > 7$ and $dos > 0.7$) for SB (blue dots) and normal papers (green dots) in the real network. The difference is much more clear, meaning, as we said, that SBs tend to be close to each other

In considering triangles we have to remember that in direct networks the position of edges in triangles is of fundamental importance to understand their role. For example, taking a look at fig 65, the edge is in position $B \to C$ was published before both the others. and so this first citation to C is independent of the other two. To understand a little bit more what

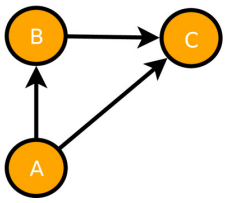

Figure 65: Schematization of a triangle

happens to triangles in our simulation, for each edge in the real and simulated network I have calculated the number of triangles in which it is in position $A \to B$, $B \to C$, $A \to C$, and compared the distributions in figures 66 68 67. In fig 69 are the three distributions of the three positions only in the simulated case.

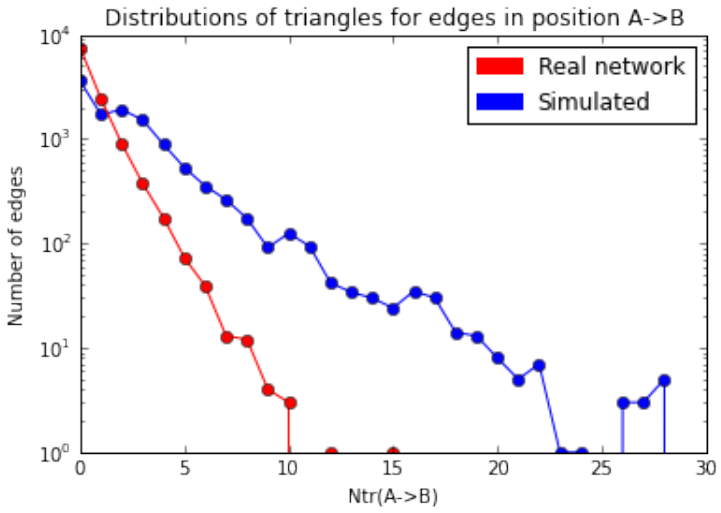

Figure 66: Distribution of triangles in position AB. y scale is logarithmic

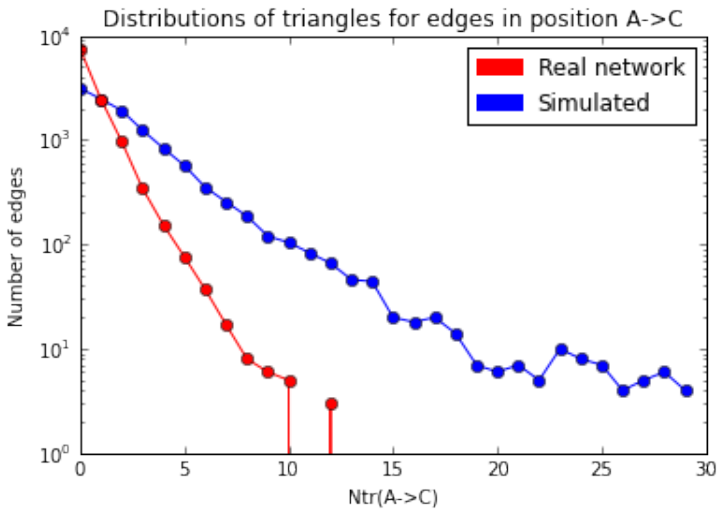

Figure 67: Distribution of triangles in position AC. y scale is logarithmic

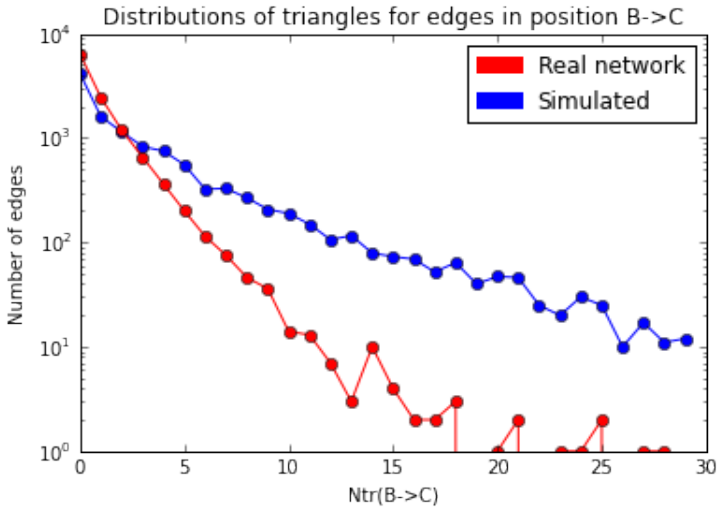

Figure 68: Distribution of triangles in position BC. y scale is logarithmic

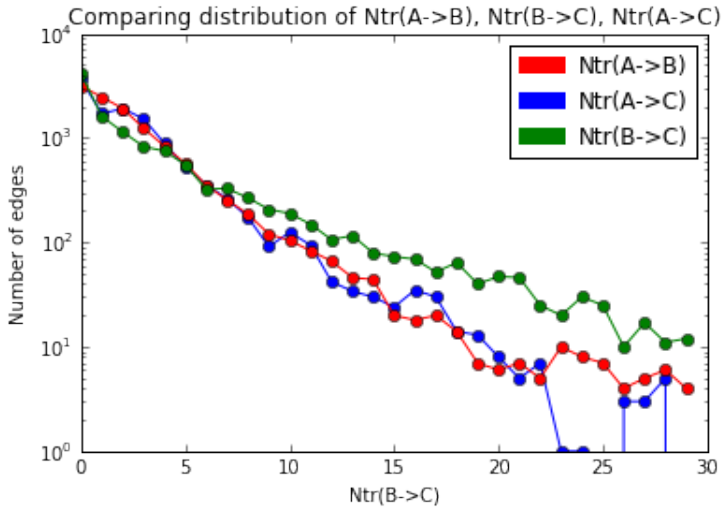

Figure 69: Comparison of the three distributions. y scale is logarithmic. As one can easily notice, distribution of $B \to C$ is slightly different from the other two, with on average a higher number of triangles

11 Conclusion

The main aim of this thesis has been to describe the behavior of Sleeping Beauties in the context of citation dynamics.
First of all, I've described some of the most important models that have been developed recently, putting an emphasis on the common features between them: the attention for time distributions and triangles, cumulative advantage and aging of papers.
Then I have introduced the concept of 'delayed recognition', starting from the definition of Sleeping Beauty given by Van Raan in 2004. Since these phenomena, from a first and superficial point of view, seem to be rare and exceptional, particularly interesting is to understand how they are generated. But in order to do so, first we have to find a good and fast definition to use in large data sets: I've focused on trying to find this algorithm, introducing the SBc and other coefficients as *awt* and *dos*, necessary to have a deeper insight in a paper's history without visualizing its history.
Then I've also tried to formulate some hypothesis on the mechanism that may trigger their appearance. One of the post popular involves, for each SB, the existence of a paper, called the prince, that gives visibility to others, inducing citations to them to the point of being alone responsible for their new life.

This 'prince' article is supposed to have some specific topological properties that involve its relationship with the SB: for example,the citing lives of these two articles have to be very similar immediately after the awakening. However, this is only rarely true: in a lot of cases it seems that independently a lot of authors suddenly decided to cite the SB.
Analyzing the network, we have also found that SBs tend to be significantly close to each other, and this does not seem to be explainable just by means of actual models since they focus only on time and triangles. Trying to be more precise on explaining why in our opinion actual fail in doing so, I've analyzed three clusters of delayed impact papers (as found by an algorithm). The presence of a prince would justify the effort of focusing only on triangles, but only in one case I could recognize such a paper for all of them. In one only two of six papers had a common prince, and in the third case there was no prince at all!
Trying to have a more systematic insight on the system, I have simulated a network preserving time and the total number of triangles: what is found is that, even with very clustered networks (by choice of the matrix of trial moves), SBs do not tend to be closer to each other than normal papers, and all the most important topological distribution of the real network appear very different in the simulated one.
This leads to the conclusion that triangles and time distribution alone can not be taken as unique ingredients for a model that aims to describe SBs and their behavior.

12 Appendices

12.1 The Theory of branching processes

The theory of branching processes was conceived in th 19th century in England: at the time, some gentlemen had noticed how some of the most powerful and influential families of the past had become extinct.
These men had concluded that an increase in intellectual capacity is accompanied by a decrease in fertility, but with the development of branching processes it was demonstrated that a large fraction of families (more specifically surnames) can become extinct just by chance.
We start from considering that in each generation an individual has a $p(0)$ probability to have no sons, $p(1)$ to have one and so on. We will use the method of generating function in order to easily get to analytic results:

$$f(z) = \sum_{n=0}^{\infty} p(n) z^n$$

The most interesting property of it is that we can easily get the generating function of the grandsons simply by combination:

$$f_2(z) = f(f(z))$$

This can be proved if we consider two individuals instead of one: since the two offspring are independent, starting from $(f(z)^2)$, the n-th term of this expansion is obviously

$$\sum_{m=0}^{\infty} p(n-m)p(m)$$

which is indeed the probability that the combined (independent) offspring of two people is n.
The same argument is true for the combined offspring of n generic people.
So the generating function for the number of grandsons is:

$$f_2(z) = \sum_{n=0}^{\infty} p(n)(f(z))^n = f(f(z))$$

and more generally

$$f_k(z) = f(f_{k-1}(z))$$

The probability of extinction of a family can be found using the self-consistency equation:

$$p_{ext} = \sum_{n=0}^{\infty} p(n) p_{ext}^n = f(p_{ext}) \tag{12}$$

A very important value that determines the fate of a family is the average number of sons:

$$\lambda = \sum_n np(n) = [f'(z)]_{z=1}$$

When $\lambda < 1$, equation 12 has only the solution $p_{ext} = 1$, i.e. every family becomes extinct.
When $\lambda > 1$, however, there is a solution where $p_{ext} < 1$, and only some of the families become extinct.
In the intermediate case of $\lambda = 1$ all the families become extinct, but some only after a very long time.

In the sub critical branching process, although the probability of extinction is still one, we can consider the probability of extinction after k generations, $p_{ext}(k)$.
Obviously:

$$p_{ext}(k) = f_k(0)$$

Considering that $p_{ext} = 1$ for large k, $p_{ext}(k)$ must be close to 1. Therefore:

$$f_k(0) = f(1) + f'(1)(f_{k-1}(0) - 1) + \frac{f''(1)}{2}(f_{k-1}(0) - 1)^2$$

Considering that $f(1) = 1$ and $f'(1) = \lambda$, if we define the survival probability as $p_s(k) = 1 - p_{ext}(k)$, we get the equation:

$$\frac{p_s(k)}{p_s(k-1)} = \lambda - \frac{f''(1)}{2}\, p_s(k-1) \tag{13}$$

First, we consider the case with $\lambda = 1$.
We can rewrite the equation as:

$$\frac{dp_s(k)}{dk} = -\frac{f''(1)}{2}(p_s(k))^2$$

that leads to the solution:

$$p_s(k) = \frac{2}{f''(1)\ k}$$

When λ is substantially less than 1, the second term in eq 13 is negligible in the approximation of large k, and we get:

$$p_s(n) \sim \lambda^k$$

Another very important estimate is the average size s(k) of the families still surviving after k generations.

The expectation value of the offspring after k generations is obviously λ^k, and so we have:

$$s(k) = \frac{\lambda^k}{p_s(k)}$$

Again, in the case of $\lambda = 1$, we have, substituting p_s into the above equation:

$$s(k) \sim \frac{f''(1)}{2}\ k$$

In the other case we have analyzed (λ much less than 1), the average size of a surviving family approaches the fixed value:

$$s(\infty) \sim \frac{f''(1)}{2(1-\lambda)}$$

For sub critical processes, we are also interested in the probability distribution of the total offspring, i.e. the sum of sons, grandsons, great grand sons, etc.
We introduce another generating function:

$$g(z) = \sum_{n=1}^{\infty} P(n) z^n$$

We can add a new self consistency condition of the form:

$$z f(g) = g$$

and using Lagrange expansion on the left term, we get to the result:

$$P(n) = \frac{1}{n!} \left[\frac{d^{n-1}}{d\omega^{n-1}} (f(\omega))^n \right]_{\omega=0}$$

Let us now consider the case in which the probability of the first generation is different from the one of the others, i.e. $\lambda_0 \neq \lambda$.
It is possible to show that the generating function for the total offspring is:

$$\tilde{g}(z) = z f_0(g(z)) \tag{14}$$

From the equation seen above, we have that:

$$f_0(z) = (f(z))^{\lambda_0/\lambda}$$

and substituting into eq 14 we get to:

$$\tilde{g}(z) = z \left(\frac{g(z)}{z} \right)^{\lambda_0/\lambda}$$

This formula can be of some use in the cases in which λ_0/λ is an integer. In our case, since

$$\frac{\lambda_0}{\lambda} = \frac{\alpha}{1-\alpha} N_{ref}$$

with $\alpha \sim 0.1$ and $N_{ref} \sim 20$, we have $\lambda_0/\lambda \sim 2$.
This way, we have:

$$\tilde{g}(z) = \left(\frac{g(z)}{z}\right)^2 = z \left(\sum_{n=1}^{\infty} P(n) z^{n-1}\right)^2$$

and so, the citation probability distribution becomes

$$\tilde{P}(n) = \sum_{l=1}^{n} P(l) P(n-l+1)$$

From this formula, we can easily get the large n-asymptotic of $\tilde{P}(n)$:

$$\tilde{P}(n) \propto 2P(n) \sum_{l=1}^{\infty} P(l) = 2\ P(n)$$

So the result is that having a different first generation offspring mean does not change the shape of the probability, but only modifies the numerical prefactor

12.2 Markov Chain Monte Carlo Methods

In broad terms, Markov chains are stochastic processes that satisfy the Markov property: the future is independent of the past given the present.
With more specificity, a Markov chain is a process of the form X_1, X_2, X_3,... with each X_i taking an arbitrary value on a given state space S, and having the property that the conditional probability:

$$P(X_{n+1}|X_n, ..., X_0) = P(X_{n+1}|X_n)$$

is independent of the past once we have conditioned on the present.
Markov chains have a lot of properties regarding their development in time, and this can be used for sampling states according to specific probability distributions.
Consider the space S and a function depending on a variable in S.
Let's suppose we want to find the points (it could be a whole manifold) that correspond to minimums of this function, but, due to some intrinsic complexity of the calculation, we are unable do it analytically. A way to get an approximate solution is to use Markov Chains, and the method is called Markov Chain Monte Carlo (MCMC in the following).
The idea is to sample the random configurations (i.e. of the minimums of the function) by performing a stochastic path on the configuration space S.
The underlined stochastic dynamics will be a Markov chain that, after a transient, reaches a steady state in which the random configurations are sampled according to a desired probability distribution π.
Instead of building them anew, this procedure speeds up the generation of configurations, since the one at time i is a deformation of the one at time $i-1$, but has also the defect that the states will be correlated. Anyway, this correlation can be made negligible just by making enough MCMC iterations.
So, the problem of sampling states from a given distribution has now reduced to finding an efficient Markov chain that has a satisfying stationary state.
In order to have a better control on what happens, we have to take a closer look to the properties of Markov Chains.

12.2.1 Properties of Markov Chains

In order to treat MC analytically, it is very convenient to introduce a transition matrix.
Let's consider a discrete state space S, and be i and j two generic states.
Be $P(X_{n+1} = j|X_n = i)$ the probability that if the system is in configuration i at time n, at time $n+1$ the system is in j. Then, if we consider only

Chains whose jump probabilities do not depend on time, we will define the transition matrix (independent of time) as:

$$\Pi(i,j) = P(X_{n+1} = j | X_n = i)$$

Then, the theory of Markov chains tell us that if we want to generate configurations X_i in a space S distributed with a given probability distribution π, we have to build up a transition matrix(TM) $\Pi(i,j)$ that satisfies:

- *irreducibility* and *aperiodicity*: A TM is said to be irreducible if for any given couple of states i, j there is a possible (with non zero total probability) path that starts in i and arrives in j. For periodicity, we define the period of a state i as:

  $$gcd\{n > 0 : P(X_n = i | X_0 = i) > 0\}$$

  where *gcd* is the greatest common divisor. If $k = 1$, then i is said to be aperiodic. If all the states in S are aperiodic, the chain (and the TM) is said to be aperiodic.
  It can be demonstrated that a state i is aperiodic if there exists n such that for every $n' > n$ we have:

  $$P(X_{n'} = i | X_0 = i) > 0$$

- *Stationarity of* π: for each $i \in S$ we must have that:

  $$\sum_{j \in S} \pi_j \Pi_{ji} = \pi_i$$

- *Detailed Balance*: for each pair of states $i, j \in S$ we must have

  $$\pi_j \Pi_{ji} = \pi_i \Pi_{ij}$$

  A MC that satisfies this condition is said to be *reversible* since $P(X_t = i, X_{t+1} = j) = P(X_t = j, X_{t+1} = i)$ at equilibrium

If all these requests are satisfied, the ergodic theorem applies, who says that the chain will reach the stationary distribution (in our case π) independently of the starting point.

Metropolis Algorithm

While the first condition must be proved for each single case, the second and the third are easily satisfied if we consider the *Metropolis filter* (note that the second condition is automatically verified if the third is).
In order to see how this filter works, we consider the matrix Φ, called the matrix of trial moves. Its role is to generate the set of proposed moves $i \to j$, that will be accepted or rejected according to a probability matrix a_{ij}. Note that the matrix Π can be easily calculated from Φ and a:

$$i \neq j \qquad \Pi_{ij} = \Phi_{ij}\ a_{ij}$$

$$i = j \qquad \Pi_{ij} = \Pi_{ii} = 1 - \sum_{j \neq i} \Pi_{ij} = \Phi_{ij} + \sum_{i \neq j} \Phi_{ij}(1 - a_{ij})$$

and so the condition $\pi_j \Pi_{ji} = \pi_i \Pi_{ij}$ becomes:

$$\pi_i \Phi_{ij} a_{ij} = \pi_j \Phi_{ji} a_{ji}$$

that we can also write, if $i \neq j$, as:

$$\frac{a_{ij}}{a_{ji}} = \frac{\pi_j \Phi_{ji}}{\pi_i \Phi_{ij}}$$

and in order to fulfill this request, it is sufficient that

$$a_{ij} = F\left(\frac{\pi_j \Phi_{ji}}{\pi_i \Phi_{ij}}\right)$$

where F is a function that takes values only in the interval $[0, 1]$, and that satisfies the condition:

$$\frac{F(z)}{F(1/z)} = z$$

There are two very popular choices for F

- *Metropolis*: $F(z) = min(z, 1)$
- *Heat Bath*: $F(z) = \frac{z}{1+z}$

In the case of statistical mechanics, each state in S is distributed according to the canonical probability distribution:

$$\pi_i = \frac{1}{Z} e^{-\beta H(i)}$$

and if we want to use the Metropolis choice for F, we have that:

$$F(e^{-\beta(E(j)-E(i))})$$

and

$$a_{ij} = min(e^{-\beta\Delta E}, 1)$$

To conclude, a possible algorithm to generate states according to the probability π in the state space S is:

- from a state i at time t search for a state j using the probability distribution Φ_{ij}
- if E_i and E_j are the energies of the old and new configurations, if $E_j < E_i$ the move is accepted and j becomes the new state of the system
- if $E_j > E_i$, consider a number taken from a uniform distribution in $[0, 1]$. If this number is less than $e^{-\beta\Delta E}$, the move is still accepted, otherwise it is rejected and the system stays in i at time $t + 1$

Example: MC for Ising Model

If we consider N spins on a lattice, the energy associated to a given configuration $\sigma = \{\sigma_1, ..., \sigma_N\}$ is given by the Hamiltonian:

$$H = \sum_{\langle i,j \rangle} \sigma_i \sigma_j$$

where $\sigma_i = \pm 1$ and $\langle i, j \rangle$ means that i and j are nearest neighbors.
Obviously, the presence of an external magnetic field that couples to all the spins linearly would break the rotational symmetry. However, when the temperature of the system is cooled under a certain T_c, the symmetry is again broken (spontaneously).
Ising solved the model in one dimension analytically and showed that there is no phase transition for any values of T different from 0, but the situation is completely different when we add dimensions to the system. On the other hand, in $d = 2$ calculation get easily very complicated, and one needs to use approximate computational methods.
In order to use MCMC method, first we need to choose the matrix of trial moves Φ. A simple choice could be the one that has as single option the flip of the spin: from configuration $i = \{\sigma_1, ..., \sigma_k, ..., \sigma_N\}$ to $j = \{\sigma_1, ..., -\sigma_k, ..., \sigma_N\}$.
This means to use:

$$\Phi_{ij}(k) = 1$$

if i and j are the ones above,

$$\Phi_{ij}(k) = 0$$

for any other choice. Moreover, $\Phi(k)$ is irreducible and aperiodic, and we can use Metropolis filter:

$$a_{ij}(k) = min(e^{-\beta \Delta E}, 1)$$

with

$$\Delta E = H(i) - H(j) = 2\sigma(k) \sum_{l} \sigma_l$$

where we have put $J = 1$ for simplicity.
k can be chosen with any specific spanning: we will simply pick it out by random from the locations in the lattice. After creating a 2 dimensional array, and having initialized it (for every location, picked a random number between 0 and 1. If more than 0.5 spin up, otherwise spin down), we can start with the algorithm as described in the previous section. Before sampling, we have to wait some time until the system gets to the stationary situation in which we get states from the canonical distribution (Fig 70)

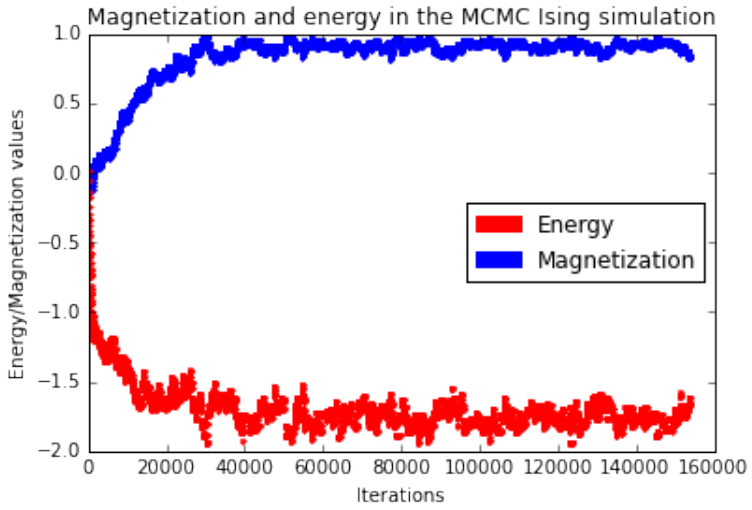

Figure 70: This graph shows how M and E develop during the iteration of MCMC algorithm with a lattice of 400 sites, J=1 and T=2 (H=0). It takes almost 40 000 iterations before starting to sample properly from the canonical distribution.

.

Another non trivial problem is that if the systems finds a stationary state, it stays in that situation for a long time as if it was actually in equilibrium.

A way to notice this unlucky and unwanted situation is to let the system work for enough time even after reaching any stationary situation before starting to sample from it.
Finally, as already anticipated, we know that with this method the states that we sample are correlated with one another, but we also know that this correlation (for the properties of Markov chains) decreases exponentially with the iterations. What we expect is that:

$$\chi(t) \sim e^{t/\tau_{int}}$$

in order to have a good independence of the states, normally one has to wait for a time interval of 2τ.
We will calculate $\chi(t)$ (i.e. the correlation function for M) with the approximate formula:

$$\chi(t) = \frac{1}{t_{max}-t}\sum_{t'=0}^{t_{max}-t} m(t')m(t'+t) - \frac{1}{t_{max}-t}\left(\sum_{t'=0}^{t_{max}-t} m(t')\right)\frac{1}{t_{max}-t}\left(\sum_{t'=0}^{t_{max}-t} m(t'+t)\right)$$

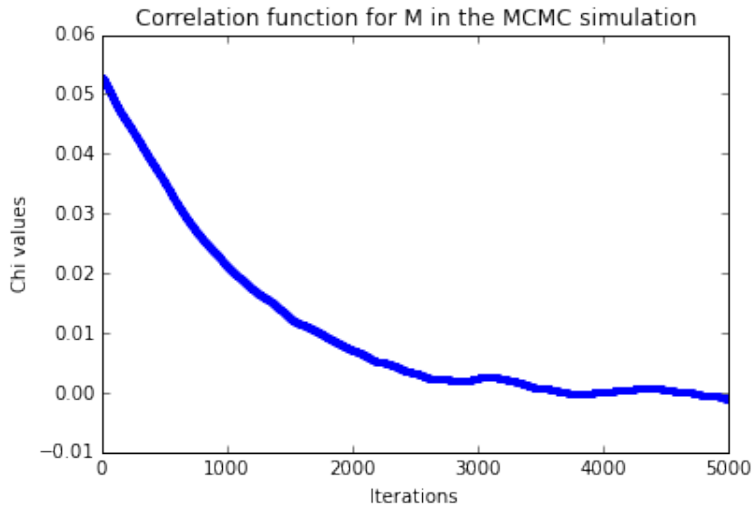

Figure 71: Correlation for M in the states sampled using the MCMC algorithm. As it's easily seen, only after 3000 iterations two states can be considered independent. Before that, values of M are very dependent one from the other.

.

Using this formula, one can calculate the correlation of M in the states sampled using MCMC, and the result is shown in Fig 71

References

[1] Moreno, J. L., *Who Shall Survive?*, New York, N.Y.: Beacon House, 1934.

[2] Erdos, Renyi, A. *On Random Graphs* Publicationes Mathematicae 6: 290-297, 1959.

[3] Park, Newman, *The statistical mechanics of networks*, arXiv, 2004

[4] H. A. Simon, *On a class of skew distribution functions*, Biometrika 42, 425-440, 1955

[5] Albert-Laszlo Barabasi, Reka Albert, *Emergence of scaling in random networks*, Science, 1999.

[6] Mark Newman, *The structure of scientific collaboration networks*, PNAS, 2001.

[7] Derek de Solla Price, *Network of scientific papers*, Science, 1965.

[8] Mark Newman, *The first-mover advantage in scientific publication*, arXiv, 2008.

[9] Brian karrer, Mark Newman, *Random graph models for directed acyclic networks*, arXiv, 2009.

[10] Zhi-Xi Wu, Petter Holme, *Modeling scientific-citation patterns and other triangle-rich acyclic networks*, arXiv, 2009.

[11] Simkin, Roychowdhury, *Read Before You Cite!*, Complex Systems Publications, 2003.

[12] Simkin, Roychowdhury, *A Mathematical Theory of Citing*, American Journal for information science and technology, 2007.

[13] Michael Golosovsky, Sorin Solomon, *Uncovering the dynamics of citations of scientific papers*, arXiv, 2014.

[14] Dashun Wang, Chaoming Song, Albert-Laszlo Barabasi *Quantifying Long-Term Scientific Impact*, Science, 2013.

[15] Rolf Ulrich, Jeff Miller, *Information processing models generating lognormally distributed reaction times*, Journal of mathematical psychology, 1993.

[16] Anthony van Raan, *Sleeping Beauties in science*, Scientometrics, 2004.

[17] Tibor Braun, Wolfgang Glanzel, Andras Schubert, *On Sleeping Beauties, Princes and other tales of citation distributions ...*, Research Evaluation, 2010.

[18] Qing Ke, Emilio Ferrara, Filippo Radicchi, Alessandro Flammini, *Defining and Identifying Sleeping Beauties in Science*, PNAS, 2015.