UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics

Final Dissertation

Analysis and Modelling of Information Ecosystems

Thesis supervisor

Prof. Samir Suweis

Thesis co-supervisor

Dr. Caterina De Bacco

Thesis co-supervisor

Dr. Carlos A. Plata

Candidate

Emanuele Pigani

Academic Year 2018/2019

Contents

CHAPTER 1

INTRODUCTION

One of the most fascinating aspects of physics is the simplicity of its laws. Almost all the fundamental equations can be written in just a few lines: Maxwell's equations for classical electrodynamics, Schroedinger's equation for quantum mechanics and Einstein's equations in general relativity are only three among all possible examples. We scientists have indeed a simple and optimistic conception of the world: nature is lawful, and the same basic laws hold everywhere and forever. Everything is simple, even if not easy, and expressible in mathematical terms.

Nevertheless, if the laws are so simple, why the world outside physics classrooms seems to possess such an amazing complexity? We observe a complex reality at very different scales and with a diversity of origins. Complexity is observed at all scales: from huge mountains to the salt spray coming off a wave, from natural ecosystems formed by communities of living creatures to the interactions between people on social networks. Nature is capable of obeying simple laws, and at the same time generating complex structures. For these reasons, in the last decades a novel branch of Physics has been established: the so called Physics of Complex Systems aimed at understanding how complexity can emerge from local and simple dynamics.

1.1 COMPLEXITY

Although the ubiquity of complex systems, the concept of complexity is not easy to define rigorously. As pointed out by the Nobel Prize winner Murray Gell-Mann [27], a great number of quantities have been proposed as measures of complexity. In fact, a variety of different measures would be required to capture all our intuitive ideas about what is meant by complexity and by its opposite, simplicity.

When applied to entities in the real world, all such quantities become to some extent context-dependent or even subjective, because they depend on the level of detail of the

description of the entity, on the previous knowledge and understanding of the world that is assumed, on the language employed, and also on the method for coding the information. Furthermore, many of these quantities are uncomputable.

We do not aim herein at solving the technical difficulties of defining a rigorous observable to measure complexity, but at giving a simple intuition of complexity. In the following, we will say that a system is complex when this is composed by many interacting components which leads to the global system to exhibit some emerging patterns and or collective behavior. Our definition of complexity entails that statistical physics, which was funded to connect the macroscopic behaviour of physical systems to their microscopic description, will provide the fundamental tools to address our problems. Specifically, this thesis is devoted to the analysis of emergent patterns in social networks (e.g. Twitter, You Tube).

1.2 Statistical Physics of Social Dynamics

An online social network (OSN) is a virtual social structure made of individuals that use the Internet as a communication medium for interacting, sharing contents and opinions.

Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content. They provide access to a very vast source of information on an unprecedented scale. In particular, online social networks play a major role in the diffusion of information by increasing the spread of novel information and diverse viewpoints [8]. They have proved to be very powerful in many situations, like Facebook during the 2010 Arab spring [30] or Twitter during the 2008 U.S. presidential elections [32] for instance.

Therefore, online social networks constitute nowadays mainstream communication channels to interact, exchange opinions, and reach consensus. The abundance of information to which users are exposed through online social networks is so huge that it exceeds the capacity to consume it, leading to a competition among both memes and users for either individual or collective attention. Aspects of competition for limited attention have been studied through news, movies, and topics posted on blogs and social media, showing that this competition significantly shapes the topology and the dynamics on these information-driven platforms: users thrive for visibility, while memes resemble entities that compete for users' attention.

Nevertheless, even in simple examples, it is hard to disentangle the effects of limited attention from many concurrent factors, such as the structure of the underlying social network [35, 56], the activity of users [4], the different degrees of influence of information spreaders [47], the intrinsic quality of the information they spread [8], and the persistence of topics [46, 59]. In fact, due to the complexity of these factors, there is still no well-established theory for online social networks and at this point several aspects remain hidden. First, the understanding of how or why structural patterns emerge, change and disappear in time. Furthermore, we do not have insights on how these macro (system-wide) patterns affect (or are affected by) the dynamics of the interacting elements (users and memes) that form the network. Therefore, we

stand merely at a descriptive level. Abundance of available data let us easily describe observed macroscopic behaviours, but give no theoretical framework for predicting such patterns.

Still a sistematic results found in data is that even in the case when very dissimilar individuals (in terms of popularity, activity, but also age, etc..) interact with the OSN, some macroscopic behaviours and stunning global regularities—like the emergence of consensus about a specific issue, the distributoin of popularity, etc.—do exists. These macroscopic phenomena naturally call for a statistical physics approach to social behaviour, i.e., the attempt to understand regularities at large scale as collective effects of the interaction among single individuals, considered as relatively simple entities. The focus of the statistical physics approach to social dynamics is then to understand how these regularities and universality come about.

A conceptual difficulty may be immediately pointed out when trying to approach social dynamics from the point of view of statistical physics. In many physical situations, such as a gas, the elementary components are relatively simple objects, like atoms and molecules, whose behaviour is very well known: the macroscopic phenomena are not due to a complex behaviour of single entities, rather to nontrivial collective effects resulting from the interaction of a large number of quite simple elements. On the opposite, humans are far from being simple entities: the detailed behaviour of each of them is already the complex outcome of many physiological and psychological processes, still largely unknown. Moreover, even if one knew the very nature of such dynamics and such processes, they would be much more complicated than, say, the forces that atoms exert on each other. It would be impossible to describe them precisely with simple laws and few parameters.

The critique that models used by physicists to describe social systems are too simplified to describe any real situation is most of the times very well grounded. This applies also to highly acclaimed models introduced by social scientists, as Schelling's model for urban segregation [48] and Axelrod's model [5] for cultural dissemination. But in this respect, statistical physics brings an important added value. In most situations qualitative (and even some quantitative) properties of large scale phenomena do not depend on the microscopic details of the process. Only higher level features, as symmetries, dimensionality or conservation laws, are relevant for the global behaviour. With this concept of universality in mind one can then approach the modelling of social systems, trying to include only the simplest and most important features of single individuals and of their interactions, looking for qualitative features exhibited by models.

Beside statistical physics, another conceptual framework and related methodological tools and approaches that can be useful to study OSN, come from theoeretical ecology. In fact, competitive and cooperative interactions in online social networks are reminiscent of those present in natural ecological systems, where species interact in different ways to exploit resources and obtain food. Specifically, one can identify some parallelism between the competition between species for resources in the natural environment and how users and memes compete for attention. The introduction of a networked perspective in ecological modelling since the early 2000s (e.g. [10]) represented a major conceptual leap, establishing the link between ecosystem dynamics

and species interaction patterns. In particular, the emergence of ubiquitous structural arrangement, like nestedness [2], has been observed in many different ecosystems, which has ignited several research in order to investigate the relation between the properties of the species network and the system's stability and biodiversity [11, 50]. Notwithstanding, other structural arrangements such as modularity [9] or in-block nestedness [49] have been studied.

There have been some efforts to drive such a qualitative resemblance between natural and information ecosystems towards more quantitative exploration. In this direction, Borge-Holthoefer et al. [15] have found evidence of nested structural signatures - a landmark feature in the time evolution of natural mutualistic systems - resolved online communication discussions. As a matter of fact, nestedness in these empirical data was found alongside other architectures (modularity [9, 25], in-block nestedness [49]), depending on the maturity (time since the onset) of the online discussion under study.

All these suggestive findings (and associated gaps) point out that there is ample room in ecology in this new informational context. In the present work, we aim at exporting some successful and well-established techniques used in ecology to social interactions, and, in particular, those related to the so called neutral theory of ecology (cf. 2.4).

While modelling information ecosystems is hardly in its infancy, a statistical mechanics of ecological systems, the conceptual and mathematical framework to relate the macro and micro structural levels have been developed in ecology in the last decade and can be exploited in this context [6]. The advantage of such an approach is twofold. On the one hand, all the theoretical predictions arising from the neutral theory for ecosystems could be adapted and used for online social networks. On the other hand, a very appealing feature of social networks is the possibility of collect data easily. Unlike in the natural environment, where it is typically very hard to collect data—e.g. counting the number of individuals inside an island or of different plants in a forest—especially for long time periods. Many data coming from information ecosystems are instead very easily available and with high temporal frequency, simplifying the comparison between theoretical results and the real data.

1.3 A HINT OF NETWORK THEORY

One of the most important aspects always present in social dynamics is topology, i.e., the structure of the interaction network describing who is interacting with whom, how frequently and with which intensity. This structures can be soo complex and complicated (see e.g., the representation of Internet networks in Fig. 1.1), that a theory of networks is necessary for efficiently describing their properties. Far from aiming at extensively analysing all the existing network models, in this section, we review the main ideas and concepts of Network Theory that will be exploited and used in this work.

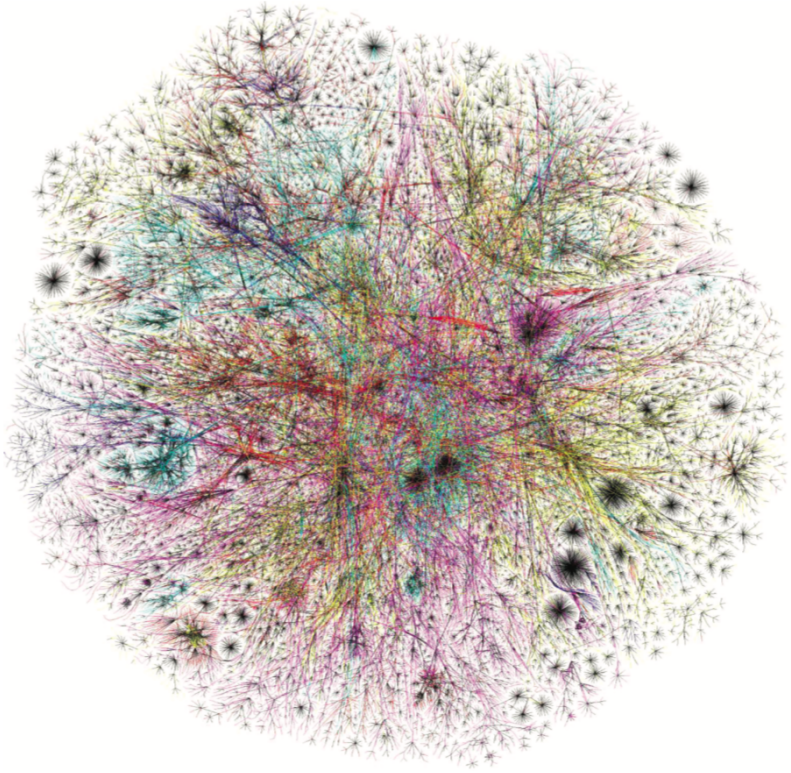


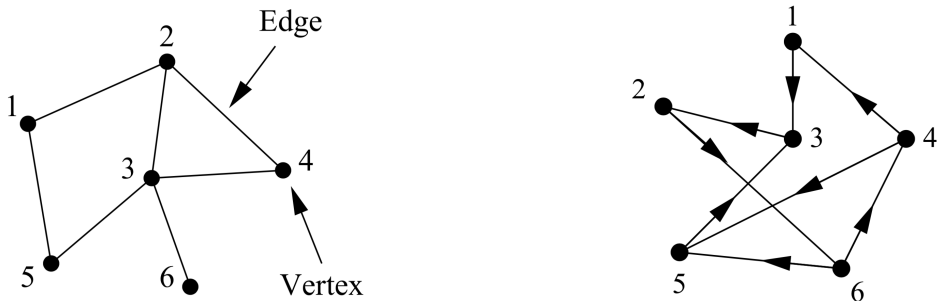
FIGURE 1.1: A pictorical representation of the Internet network. The vertices in this representation of the Internet are "class C subnets"—groups of computers with similar Internet addresses that are usually under the management of a single organisation—and the connections between them represent the routes taken by Internet data packets as they hop between subnets. Credits [41].

It is very common to describe the topology in social dynamics and ecosystems, as well in many other fields, by mean of networks or graphs, in which individuals and their possible interactions are encoded as nodes (or vertices), and edges (or links), respectively.

More specifically, an undirected **graph** G is a tuple (V, E), where the set of vertices V is identified by the N elements of the system, known as *order* of G, and the set of edges E is identified by all the M relations among them, known as *size* of G. Each edge

is an unordered pair of vertices. The two vertices associated with an edge e are called the end-vertices of e.

There are a number of different ways to represent a network mathematically. Consider a network with N vertices and let us label the vertices with integer labels $1, \ldots, N$, as we have, for instance, for the network in Fig. 1.2a. Of course, it does not matter which vertex gets which label, only that each label is unique, so that we can use the labels to refer to any vertex unambiguously.



(A) A simple grap. (B) A simple directed graph.

FIGURE 1.2: Two examples of graphs.

If we denote an edge between vertices i and j by (i, j) then the complete network can be specified by giving the value of N and a list of all the edges. A usually better representation of a network is the adjacency matrix. The adjacency matrix A of a simple graph is the (symmetric) matrix with elements A_{ij} such that

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge between nodes } i \text{ and } j, \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

For example, the adjacency matrix of the network in Fig. 1.2a is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{1.2}$$

In many situations of great interest, the relationship between nodes is not mutual, but directional. Hence, in cases like those presented in Chapters 4 and 3, it is useful to introduce a **directed graph** in which each edge has a direction, pointing from one vertex to another. The adjacency matrix in the case of a directed network is in general non symmetric and has matrix elements

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge pointing from } j \text{ and } i, \\ 0, & \text{otherwise.} \end{cases} \tag{1.3}$$

As an example, the adjacency matrix of the small network in Fig. 1.2b reads

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1.4}$$

1.3.1 DEGREE

The **degree** of a vertex in a graph is the number of edges connected to it. We denote the degree of vertex i by k_i. For an undirected graph of N nodes the degree can be written in terms of the adjacency matrix as

$$k_i = \sum_{j=1}^{N} A_{ij}. \tag{1.5}$$

The maximum possible number of edges in a simple graph is $\binom{N}{2}$. The *density* ρ of a graph is the fraction of these edges that are actually present:

$$\rho = \frac{\frac{1}{2}\sum_i^N k_i}{\binom{N}{2}} = \frac{\frac{1}{2}\sum_{ij} A_{ij}}{\binom{N}{2}}, \tag{1.6}$$

which lies strictly in the range $0 \leq \rho \leq 1$.

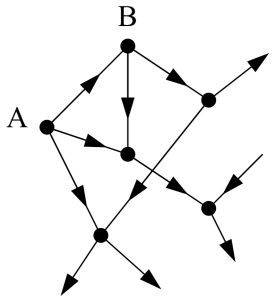
A network for which the density ρ tends to a constant as $N \to \infty$ is said to be dense. In such a network the fraction of non-zero elements in the adjacency matrix remains constant as the network becomes large. A network in which $\rho \to 0$ as $N \to \infty$ is said to be sparse, and the fraction of non-zero elements in the adjacency matrix also tends to zero.

Vertex degrees are more complicated in directed networks. In a directed network each vertex has two degrees. The **in-degree** is the number of ingoing edges connected to a vertex and the **out-degree** is the number of outgoing edges. From the adjacency matrix definition, in- and out-degrees can be written

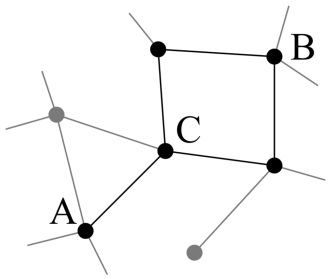
$$k_i^{in} = \sum_{j=1}^{N} A_{ij}, \quad k_i^{out} = \sum_{j=1}^{N} A_{ji}. \tag{1.7}$$

1.3.2 CENTRALITY

A large volume of research on networks has been devoted to the concept of **centrality**, which is a measure of the relevance of nodes in a network. There are of course many possible definitions of relevance, and correspondingly many centrality measures for networks.



(A) A portion of a directed network. Node A has only outgoing edges and hence will have eigenvector centrality zero. Vertex B has outgoing edges and one in-going edge, but the ingoing one originates at A, and hence vertex B will also have centrality zero.



(B) A portion of graph. Nodes A and B are connected by two geodesic paths. Vertex C lies on both paths.

FIGURE 1.3: Two examples of graphs.

1.3.2.1 Degree Centrality

The simplest centrality measure in a network is of course just the degree of a vertex. Degree is sometimes called *degree centrality* in the social networks literature, to emphasise its use as a centrality measure. In directed networks, vertices have both an in-degree and an out-degree, and both may be useful as measures of centrality in the appropriate circumstances. Although degree centrality is a simple centrality measure, it can be very illuminating. In social networks, for instance, it seems reasonable to suppose that individuals who have connections to many others might have more influence than those who have fewer connections.

1.3.2.2 Eigenvector Centrality

A natural extension of degree centrality is the so-called *eigenvector centrality*, that aims at increasing a node's relevance whenever it has connections to other nodes that are themselves important. The algorithm for computing this centrality is the following.

Let $\boldsymbol{x}(0)$ be the initial guess about the centrality, in which each vertex has the same centrality $\boldsymbol{x}_i(0) = 1$. We can use this to calculate a better centrality measure, which we define to be the sum of the centralities of i's neighbours, namely:

$$x_i(1) = \sum_j A_{ij} x_j(0). \tag{1.8}$$

Repeating this process to make better estimates, we have after t steps a vector of centralities $\boldsymbol{x}(t)$ given by $\boldsymbol{x}(t) = A^t \boldsymbol{x}(0)$. In the limit $t \to \infty$, this equation can be rewritten as an eigenvalues equation

$$A\boldsymbol{x} = \kappa_1 \boldsymbol{x} \tag{1.9}$$

where κ_1 is the largest eigenvalue of A. This is so since after reiterative application of A, in the limit $t \to \infty$, the dominant behaviour is given by the eigenvector corresponding

to the largest eigenvalue. Thus, the correct definition of eigenvector centrality, first proposed by Bonacich [**Bonacich72**] in 1972, for a node i makes it proportional to the centralities of the vertices that point to i, namely

$$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j. \tag{1.10}$$

In theory eigenvector centrality can be calculated for either undirected or directed networks. However, there are some problems with eigenvector centrality on directed networks. Consider Fig. 1.3a. Node A in this figure is connected to the rest of the network, but has only outgoing edges and no incoming ones. Such a node will always have centrality zero because there are no terms in the sum in Eq. (1.10). This might not seem to be a problem: perhaps a node that no one points to should have centrality zero. But then consider node B, which has one ingoing edge, but that edge originates at node A, and hence B also has centrality zero, because the only possible non-vanishing term in its sum in Eq. (1.10) is also zero.

1.3.2.3 PageRank Centrality

One possible solution to the gaps left by the eigenvector centrality is the *PageRank* [16]. This centrality is defined by an extension of Eq. (1.10), namely,

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta, \tag{1.11}$$

where α and β are two exogenous parameters that are independent of the network structure. This definition has a double advantage. On the one hand, the additional term β avoids the issue pointed out above, because even nodes with zero in-degree still get centrality β, and once they have a non-zero centrality, then the nodes they point to derive some advantage from being pointed to. On the other hand, dividing by k_j^{out} will dilute the contribution of node j to all the nodes that are connected with it.

This modification embodies the idea that nodes that point to many others pass only a small amount of their centrality on to each of those others, even if their own centrality is high.

1.3.2.4 Betweenness Centrality

A very different concept of centrality is *betweenness centrality*, which measures the extent to which a node lies on paths between other nodes. Suppose we have a network with something flowing around it from vertex to vertex along the edges. For instance, in a social network like Twitter we might have memes, news, or information being passed from one user to another. In this scenario, one would say that nodes with highest betweenness are the ones through which the largest number of messages pass. The nodes with highest betweenness are also the ones whose removal from the network will most disrupt communications between other nodes because they lie on the largest number of paths taken by messages. Consider for instance Fig 1.3b: if node C is removed, there is no path left between A and B.

This is the key idea of betweenness centrality, that can be formalised as folllows. Let n_{st}^i be 1 if vertex i lies on the geodesic path from s to t and 0 if it does not or if there is no such path (because s and t lie in different components of the network). Then the betweenness centrality x_i is given by

$$x_i = \sum_{st} n_{st}^i. \tag{1.12}$$

1.3.3 ASSORTATIVITY

Suppose we have a network in which the nodes are classified according to some characteristic that has a finite set of possible values. In many cases like OSN, the nodes could represent people and be classified according to many features such as nationality, political orientation, or gender. It turns out in many systems (see e.g., [39]) that people have a tendency to associate with others whom they perceive as being similar to themselves in some way. This tendency is called *homophily* or *assortative mixing*.

A network is said to be assortative if a significant fraction of the edges in the network run between nodes of the same type. The simplest way to quantify assortativity would be to measure that fraction; however, this is not a very good measure because, for instance, it is 1 if all nodes belong to the same single type. This is a trivial sort of assortativity: all friends of a human being, for example, are also human beings, but this is not really an interesting statement. What we would like instead is a measure that is large in non-trivial cases but small in trivial ones.

A good measure turns out to be the following. For the sake of simplicity, let us consider undirected graphs, the generalisation for directed graphs being straightforward. We find the fraction of edges that run between nodes of the same type, and then we subtract from that figure the fraction of such edges we would expect to find if edges were positioned at random without regard for vertex type.

Let us then denote by c_i the class or type of vertex i, which can be indexed by an integer $1, \dots, n_c$, with n_c being the total number of classes. Then the total number of edges that run between nodes of the same type is

$$\sum_{\text{edges}(i,j)} \delta(c_i, c_j) = \frac{1}{2} \sum_{ij} A_{ij} \delta(c_i, c_j), \tag{1.13}$$

where the introduced δ function equals either to (i) one if the consider edge links two nodes with the same class or to (ii) zero on the contrary. Instead, the expected number of edges between all pairs of vertices of the same type is

$$\frac{1}{2} \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j), \tag{1.14}$$

where m is the total number of edges. We therefore define the **modularity** as

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j). \tag{1.15}$$

This quantity is strictly less than 1, takes positive values if there are more edges between nodes of the same type than we would expect by chance, and negative ones if there are less.

An alternative form for measuring the assortativity is the **assortativity coefficient**, which is usually more efficient to be computed. For this purpose, consider the quantity e_{xy}, which is defined as the fraction of edges in a network that connect a node of type x to one of type y. On an undirected network this quantity is of course symmetric in its indices , i.e., $e_{xy} = e_{yx}$, while on directed networks or bipartite networks it may be asymmetric. It satisfies the sum rules

$$\sum_{xy} e_{xy} = 1, \quad \sum_{y} e_{xy} \equiv a_x, \quad \sum_{x} e_{xy} \equiv b_y, \tag{1.16}$$

where a_x and b_y are the fraction of each type of end of an edge that is attached to vertices of type x. On undirected graphs, where the ends of edges are all of the same type, $a_x = b_x$. Furthermore, if there is no assortative mixing $e_{xy} = a_x b_y$. If there is assortative mixing one can measure it by the assortativity coefficient r

$$r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}, \tag{1.17}$$

which is nothing else than the standard Pearson correlation coefficient [44], while σ_a and σ_b are the standard deviations of the distributions a_x and b_y.

CHAPTER 2

MODELS FOR MEME POPULARITY

Twitter is a social networking service on which users post and interact with messages known as "tweets". Launched in July 2006, in few years it became one of the most known and used social network. In 2013, it was one of the ten most visited websites, with more than 100 million of users [51]. As of 2018, it had more than 321 million of monthly active users [52].

The peculiarity of Twitter is that tweets are publicly visible by default. Indeed, as a social network, Twitter revolves around the principle of "followers". Users may subscribe to other users, which is known as "following", and may read all of their tweets, which appear in reverse chronological order on user's main Twitter page, as long as they are published. Users often mark their posts with topic labels ("hashtags"), which are usually used as operational proxies to identify memes[1]. Moreover, memes can be forwarded by other users to their own feed with a process known as a "retweet".

As a meme spreads in its way, it forms a cascade or diffusion network such as those illustrated in Fig. 2.1. When people select from multiple tweets of roughly equal value and retweet them, some items quickly become extremely popular, while other items are chosen by relatively few people. Let us introduce a relevant quantity that will play an important role in this work. We define the popularity n as the number of times that an item is spread. The probability $P_n(t)$ that a random item has been selected n times by time t is often observed to have a heavy-tailed distribution. Indeed, it has been observed that in examples where the tweets are URLs or hashtags 2.1, the popularity distribution is found to scale approximately as a power law $P_n \sim n^{-\alpha}$ over several decades. The exponent α in all these examples is less than two, and typically has a value

[1]Herein, we define memes as ideas which are spread from users to users. The meme can be understood as the cultural analogue to gene in that they self-replicate, mutate, and respond to selective pressures.

close to 1.5. Interestingly, the value $\alpha = 1.5$ is also found for the power-law distribution of avalanche sizes in self-organized criticality (SOC) models [7, 60], suggesting the possibility that the heavy-tailed distributions of popularity in the examples above are due to the systems being somehow poised at criticality.

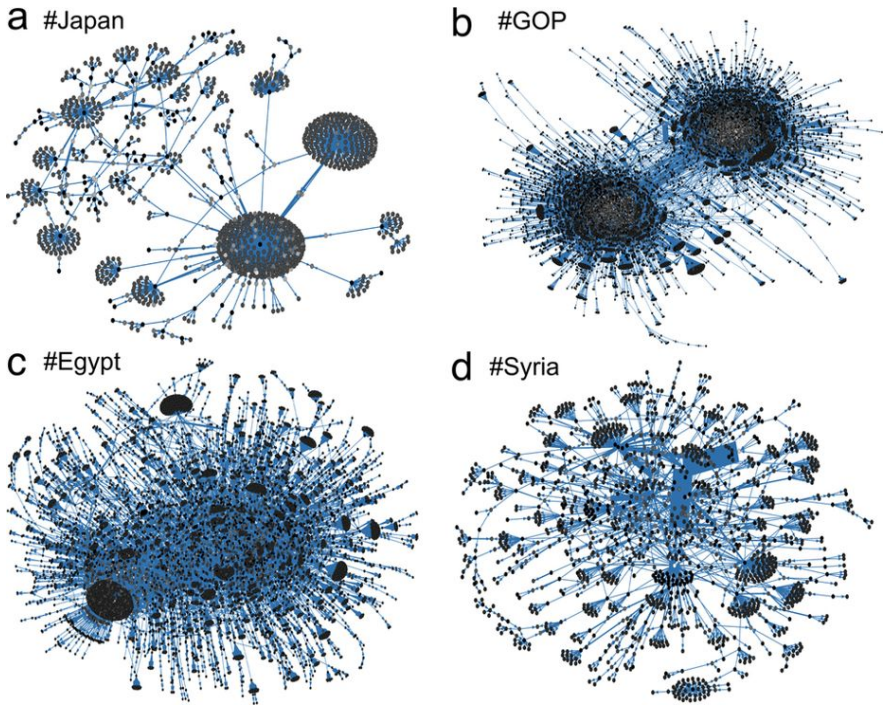


FIGURE 2.1: Nodes represent Twitter users, and directed edges represent retweeted posts that carry the meme. The brightness of a node indicates the activity (number of retweets) of a user, and the weight of an edge reflects the number of retweets between two users. (a) The Japan meme shows how news about the March 2011 earthquake propagated. (b) The GOP tag stands for the US Republican Party and as many political memes displays a strong polarisation between people with opposing views. Memes related to the Arab Spring and in particular the 2011 uprisings in (c) Egypt and (d) Syria display characteristic hub users and strong connections, respectively. Credits [57].

In this chapter we present an analytically tractable model for social interaction, proposed by J. P. Gleeson [28], for the spreading of memes on a social network. The system describes the competition between memes for the limited resource of user attention. Furthermore, we extend the study of this model by considering not only the popularity of memes, but also their exposure among users, which is a measure of their visibility. Besides that, we propose a theoretical derivation for the popularity distribution $P_n(t)$ based on path integral methods for stochastic differential equations.

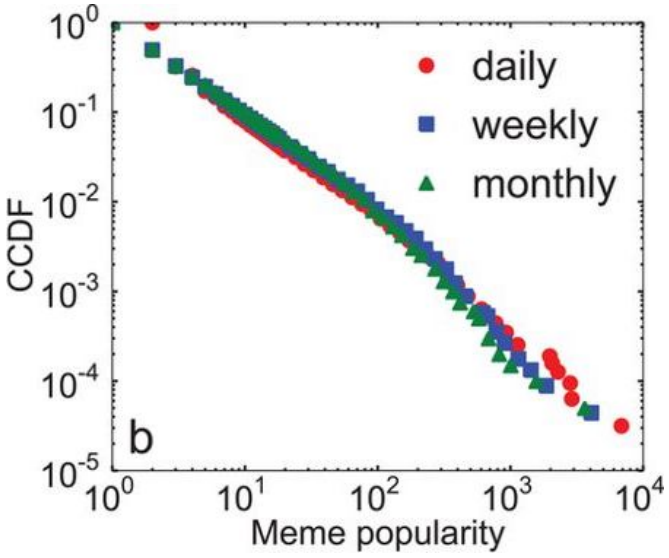


FIGURE 2.2: Complementary cumulative probability distribution of the popularity of a meme, measured by the total number of users per day who have used that meme. This measures were performed with different time resolutions: daily (filled red circles), weekly (filled blue squares), and monthly (filled green triangles). Credits [57].

2.1 COMPETITION-INDUCED CRITICALITY IN A MODEL OF MEME POPULARITY

As already mentioned in the introduction (cf. 1.2), many critical systems [13, 14, 24] can be described by the competition among items for limited resources. In this spirit, Gleeson et al. [28] proposed a model for Twitter's memes competing for user attention, showing that meme popularities are described by a critical branching process, in a mechanism called "competition-induced criticality" (CIC). This is the model that we will exploit to better understand Twitter dynamics along the whole chapter.

Let us consider a directed social networks, where nodes represent users. In our analysis we will consider the limit $L \to \infty$, where L is the total number of nodes, i.e., the system size. A randomly chosen user has k followers (i.e., out-degree k, note we use the convention that network edges are directed from nodes to their followers) with probability p_k. Each node has a screen, which holds the meme of current interest to that node. For simplicity, we assume that each screen has capacity for only one meme, though this case is easily extended (see Appendix A) for the popularity dynamics at least.

The dynamics is is introduced in discrete time steps. During each time step—with time increment $\Delta t = 1/L$—one node is chosen at random and then with probability

- μ: the selected node innovates, i.e., generates a brand-new meme that appears on its screen and is tweeted (broadcast) to all the node's followers;
- $1 - \mu$: the selected node (re)tweets the meme currently on its screen to all its followers, while its own screen remains unchanged.

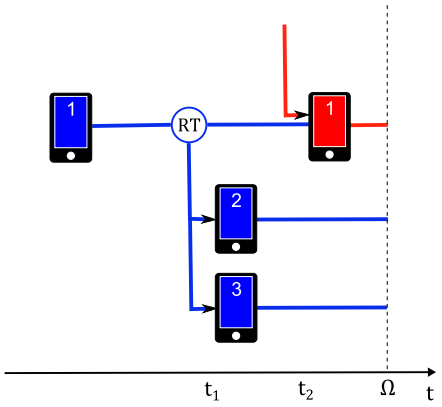


FIGURE 2.3: Schematic of the model. Time runs horizontally and nodes of the network are listed vertically; the screen colour of each node indicates the meme it currently holds. At time t_1, node 1 retweets the blue meme to its followers (nodes 2 and 3). At time t_2, node 1's screen is overwritten by the red meme, which was tweeted by one of the nodes followed by node 1. Credits [28].

As previously introduced, a quantity of great interest in our model is the *popularity* of memes: when a meme m is tweeted, its popularity is incremented by 1 and the memes currently on the followers' screens are overwritten by meme m. We denote the distribution of popularities at age a by $q_n(a)$: this is the probability that a meme has been tweeted n times when its age is a, i.e. at a time $t_b + a$, where t_b is the birth time of the meme. Note that, technically, n is always an integer number that only can increase by discrete jumps of exactly 1 when it is selected to broadcast during a time step.

The other interesting feature of the model we are interested in is the *exposure*, which is defined for every meme as the fraction x of users viewing it on their screens at a given time. Therefore, it follows that $0 \leq x \leq 1$ for all times and it can increase or decrease during its evolution.

2.2 POPULARITY DISTRIBUTION

In this section, we are interested in predicting the popularity distribution $q_n(a)$. This can be represented via its probability generating function (PGF) defined by

$$H(a, y) = \sum_{n=1}^{\infty} q_n(a)\, y^n, \tag{2.1}$$

where y is an auxiliary variable. Of course, this distribution will depend on the network topology, which could be described by the PGF for the out-degree distribution

$$f(y) = \sum_{k=0}^{\infty} p_k\, y^k, \tag{2.2}$$

where p_k is the out-degree probability distribution. In the following, we denote the mean degree as $z = f'(1)$. To calculate $q_n(a)$, we first find $H(a, y)$ and then employ an inversion technique based on fast Fourier transforms (FFTs) [18, 28].

2.2.1 PROBABILITY GENERATING FUNCTION FOR THE EXCESS POPULARITY

For computing $H(a, y)$, it proves convenient to introduce $G(a, y)$, defined as the PGF for the excess popularity distribution at age a of memes that originate from a single randomly chosen screen (the root of the tree). The tweet event that creates the root is not counted by G: this event increases the popularity of the meme by 1, and places the meme upon the root screen and the screens of all followers of the root node. It is handy to also introduce the PGF of the excess popularity distribution for a meme seeded by (i.e., first tweeted by) a node with out-degree k, $G^{(k)}(a, y)$. So, using the out-degree distribution p_k of the network, we have the relation

$$G(a, y) = \sum_k p_k G^{(k)}(a, y). \tag{2.3}$$

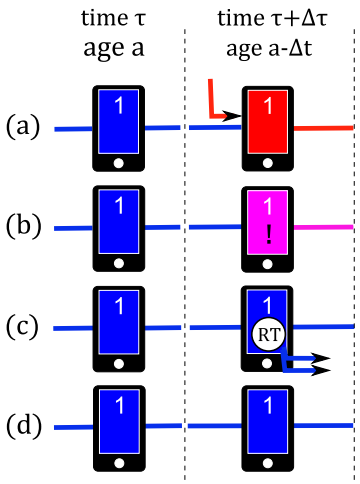
We can identify any event as the root of a retweet-cascade tree that results from the meme being tweeted and subsequently retweeted over a period of time, as shown in Figure 2.3. Deriving a relation between the PGF for the sizes of trees at age a and the PGF for tree sizes at age $a - \Delta t$ will allow us to write the evolution equation for the PGF.

Consider a meme on a given screen (let us call this screen S_1), at a given time t. This is the root of the tree we call Tree(S_1), which has age a at an observation time. Let k be the out-degree of the node with screen S_1. At the next time step $t + \Delta t$, there are four possible outcomes for this particular screen that contribute to the PGF $G^{(k)}(a, y)$, refer to Figure 2.4:

Outcome (a) the screen S_1 is overwritten by some other meme that is tweeted by another node. This terminates Tree(S_1)—setting the corresponding generating function to 1—as no future tweets can now result from the chosen root. Outcome (a) occurs with probability $z\Delta t$, since a node follows, on average, z other nodes, each of which is the tweeter with probability $1/N = \Delta t$. So outcome (a) contributes $z\Delta t$ (the probability of occurrence multiplied by the resulting generating function term) to the PGF $G^{(k)}(a, y)$ (note that all terms of order $(\Delta t)^2$ and higher are ignored here and below, as these are negligible when we take the limit $\Delta t \to 0$).

Outcome (b) the screen S_1 is selected as the updater in the current time step (with probability Δt) and innovates (with probability μ), so terminating Tree(S_1), that is, setting the PGF to 1. The total contribution to $G^{(k)}(a, y)$ is then $\mu\Delta t$.

Outcome (c) the screen S_1 is selected for update (probability Δt) and retweets its meme (probability $1 - \mu$). This (i) adds one to the size of Tree(S_1), whilst (ii) the branch on screen S_1 survives another time step, becoming the origin of a new tree rooted at $t + \Delta t$, which has age $a - \Delta t$ at the observation time. Moreover, (iii) screen S_1 is the parent of k new branches of Tree(S_1): each new branch acts as the root of a tree with PGF $G(a - \Delta t, y)$ (since the out-degrees of the newly-spawned roots are random). These effects (i) - (iii) lead to generating function contributions of y, $G^{(k)}(a - \Delta t, y)$, and $[G(a - \Delta t, y)]^k$, respectively, and since these occur simultaneously, the overall contribution of outcome (c) to $G^{(k)}(a, y)$ is $(1 - \mu)\Delta t\, y\, G^{(k)}(a - \Delta t, y)[G(a - \Delta t, y)]^k$.



	Probability	**Contribution to $\mathbf{G^{(k)}(a,y)}$**
(a)	$z\,\Delta t$	1
(b)	$\mu\,\Delta t$	1
(c)	$(1-\mu)\,\Delta t$	$yG^{(k)}(a-\Delta t,y)[G(a-\Delta t,y)]^k$
(d)	$1-(z+1)\Delta t$	$G^{(k)}(a-\Delta t,y)$

FIGURE 2.4: Summary of the single-timestep outcomes that contribute to the PGF $G^{(k)}(a,y)$.

Outcome (d) the survival of Tree(S_1), with none of the other outcomes (a) - (c) occurring: the probability of this is $1-(z\Delta t+\mu\Delta t+(1-\mu)\Delta t)=1-(z+1)\Delta t$, and the screen can then be considered as the origin of a new tree that is rooted at time $\tau+\Delta$, and so has age $a-\Delta$ at the observation time. The overall contribution to $G^{(k)}(a,y)$ from this outcome is therefore $(1-(z+1)\Delta t)G^{(k)}(a-\Delta t,y)$.

Putting all four outcomes together gives an expression for $G^{(k)}(a,y)$, correct to first order in Δt:

$$G^{(k)}(a,y)=\underbrace{z\,\Delta t}_{\text{(a)}}+\underbrace{\mu\,\Delta t}_{\text{(b)}}+\underbrace{(1-\mu)\Delta t\,y\,G^{(k)}(a-\Delta t,y)\,[G(a-\Delta t,y)]^k}_{\text{(c)}}$$

$$+\underbrace{[1-(z+1)\Delta t]G^{(k)}(a-\Delta t,y)}_{\text{(d)}}, \tag{2.4}$$

and taking the limit $\Delta t\to 0$ yields an ordinary differential equation for $G^{(k)}(a,y)$, parameterized by y:

$$\frac{\partial G^{(k)}}{\partial a}=z+\mu-(z+1)G^{(k)}+(1-\mu)y\,G^{(k)}\,[G]^k. \tag{2.5}$$

Averaging over the possible out-degrees of the root node—by multiplying by p_k and summing over all k—gives the following equation for $G(a,y)$:

$$\frac{\partial G}{\partial a}=z+\mu-(z+1)G+(1-\mu)y\sum_k p_k\,G^{(k)}\,[G]^k. \tag{2.6}$$

The equation we have obtained for $G(a,y)$ is coupled to all $G^{(k)}$, which is a challenging mathematical problem. However, if we make the following approximation (**moment-closure assumption**)

$$\sum_k p_k\left(G^{(k)}\,[G]^k\right)\approx\left(\sum_k p_k\,G^{(k)}\right)\left(\sum_k p_k\,[G]^k\right)=Gf(G), \tag{2.7}$$

which is exact in the case $p_k = \delta_{k,k_0}$ and leads to quite accurate results for networks with reasonably large mean degree z [28], we obtain the single closed differential equation for $G(a, y)$ given by

$$\frac{\partial G}{\partial a} = z + \mu - (z+1)\, G + (1-\mu)\, yGf(G). \tag{2.8}$$

This equation can be solved using standard numerical methods, starting from the initial condition $G\,(0, y) = 1$.

2.2.2 Probability Generating Function and Popularity Distribution

The Probability Generating Function $G\,(a, y)$ may be used for computing $H\,(a, y)$, since they are related by the following:

$$\begin{aligned} H\,(a,y) &= \sum_{k=0}^{\infty} p_k \cdot y\, G^{(k)}\,(a,y)\,[G\,(a,y)]^k && (2.9)\\ &\approx y \sum_{k=0}^{\infty} p_k \cdot G^{(k)}\,(a,y) \sum_{k'=0}^{\infty} p_{k'}\,[G\,(a,y)]^{k'} && (2.10)\\ &= y\, G\,(a,y)\, f\,[G\,(a,y)]\,, && (2.11) \end{aligned}$$

where again we have used moment-closure assumption given by Eq. (2.7). We notice that three factors appear in $H\,(a, y)$: y comes from the PGF for the initial seed[2]; $G^{(k)}\,(a, y)$ is the PGF for the excess popularity distribution for the initial seed with k followers, which generates itself k new trees that are taken into account by $[G\,(a, y)]^k$.

The probability $q_n(a)$ that a meme has popularity n at age a may be then determined from the PGF $H(a, y)$ by identifying the coefficients of the expansion given in (2.1),

$$q_n(a) = \frac{1}{n!}\,\frac{d^n}{dy^n} H(a,y)\bigg|_{y=0}, \tag{2.12}$$

where H is obtained using into (2.11) the (numerical) solution of Eq. (2.8). However, numerical differentiation is inaccurate for large values of n, so it is clearer to invert the PGF using contour integration in the complex y-plane [18, 42]. The inversion integral is given by Cauchy's theorem [17]

$$q_n(a) = \frac{1}{2\pi i} \oint_C H(a,y) y^{-(n+1)}\, dy, \tag{2.13}$$

where all poles of $H(a, y)$ must lie outside the contour C; a common choice for C is the unit circle [42]. Writing $y = e^{i\theta}$ gives the form

$$q_n(a) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H(a, e^{i\theta}) e^{-in\theta} d\theta. \tag{2.14}$$

Numerical integration of the above integral, involving FFT routines [18, 37], allow us to finally obtain our prediction for the popularity distribution.

[2] In contrast with G, which has an initial condition $G(0, y) = 1$, $H\,(0, y) = y$ since $H\,(0, y) = \sum_n q_n\,(0)\, y^n$ with $q_n\,(0) = \delta_{n,1}$.

2.2.3 OLD-AGE ASYMPTOTICS

The large-n asymptotic behaviour of $q_n(a)$ can be obtained in the limit $a \to \infty$ by asymptotic analysis of the solution of Eq. (2.8), with

$$\lim_{a\to\infty} H(a, y) = y\, G(\infty, y)\, f(G(\infty, y)), \tag{2.15}$$

where $G(\infty, y) = \lim_{a\to\infty} G(a, y)$ is the steady solution of Eq. (2.8), namely

$$z + \mu - (z + 1)G + (1 - \mu)y\, G\, f(G) = 0. \tag{2.16}$$

Indeed, if $G(\infty, y)$ has an asymptotic series solution for $y \to 1$, i.e.

$$G(\infty, 1 - w) \sim \sum_m c_m w^{\beta_m} \quad \text{as } w \to 0,$$

with $w = 1 - y$. We present below two lemmas that will be useful to determine the asymptotic form of q_n for large n values.

Lemma 1. *Let $\Phi(y) = \sum_{k=0}^{\infty} \pi_k y^k$ be the PGF for a well-defined probability distribution π_k, and suppose Φ has the following asymptotic series as $y \to 1$:*

$$\Phi(1 - w) \sim \textit{analytic part} + \sum_{m=1}^{\infty} c_m w^{\beta_m} \quad \textit{as } w \to 0, \tag{2.17}$$

where $w = 1 - y$ and $\beta_1 < \beta_2 < \ldots$ are positive, non-integer powers (note that the analytic part of Φ can be written as a power series in w with integer powers). Then the leading-order asymptotic behaviour of π_k is

$$\pi_k \sim \frac{c_1}{\Gamma(-\beta_1)} k^{-\beta_1 - 1} \quad \textit{as } k \to \infty, \tag{2.18}$$

where Γ stands for the Gamma function.

To prove this result, following general result from Lemma 5.3.2 of Ref. [58], we begin with the inversion integral

$$\pi_k = \frac{1}{2\pi i} \oint_C \Phi(y) y^{-k-1}\, dy, \tag{2.19}$$

where the contour C can be split into the contours $C_\epsilon \cup l_1 \cup C_R \cup l_2$ shown in Fig. 2.5. The point $y = \alpha = 1$ is a branch point, with the branch cut running from α to ∞. It is straightforward to show that the integrals along the circular contours C_ϵ and C_R go to zero as the limits $\epsilon \to 0$ and $R \to \infty$ are considered[33]. Therefore, we get

$$\pi_k = \frac{1}{2\pi i} \left[\int_{l_1} \Phi(y) y^{-k-1}\, dy + \int_{l_2} \Phi(y) y^{-k-1}\, dy \right]. \tag{2.20}$$

Along the rays l_1 and l_2, we introduce the change of variable $y = e^\rho$ to obtain integrals whose asymptotic behaviour may be determined using Watson's Lemma [12]. The

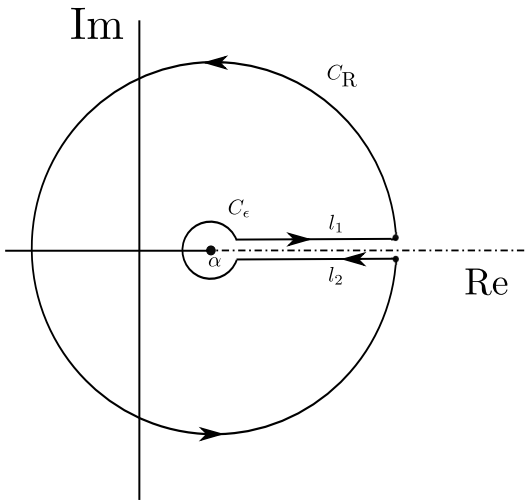


FIGURE 2.5: The contour C in the complex y-plane for the PGF inversion integral (2.19). A branch cut extends from α to ∞ (with $\alpha = 1$ for Lemma 1); the circular arcs have radii of ϵ and R.

contributions from the analytic part of Φ to the l_1 integral and to the l_2 integral cancel each other. Along the branch cut, we write the leading-order non-analytic term of (2.17) as

$$\begin{aligned} c_1 w^{\beta_1} &= c_1(1-y)^{\beta_1} \\ &= c_1(1-e^{\rho})^{\beta_1} \\ &\sim c_1(-\rho)^{\beta_1} \quad \text{as } \rho \to 0 \\ &= \begin{cases} c_1 e^{-\pi\beta_1 i}\rho^{\beta_1} & \text{as } \rho \to 0 \text{ along } l_1 \\ c_1 e^{\pi\beta_1 i}\rho^{\beta_1} & \text{as } \rho \to 0 \text{ along } l_2 \end{cases} \end{aligned}$$

to obtain

$$\begin{aligned} \pi_k &\sim \frac{1}{2\pi i}\left[\int_0^\infty c_1 e^{-\pi\beta_1 i}\rho^{\beta_1} e^{-k\rho} d\rho + \int_\infty^0 c_1 e^{\pi\beta_1 i}\rho^{\beta_1} e^{-k\rho} d\rho\right] \quad \text{as } k \to \infty \\ &= -\frac{c_1}{\pi}\sin(\pi\beta_1)\int_0^\infty \rho^{\beta_1} e^{-k\rho} d\rho. \end{aligned} \tag{2.21}$$

The integral in this expression evaluates to $\Gamma(\beta_1+1)k^{-\beta_1-1}$. Using the Euler reflection formula $\pi/\sin(\pi\beta_1) = \Gamma(\beta_1)\Gamma(1-\beta_1)$ and the Gamma function property $\Gamma(y+1) = y\Gamma(y)$ completes the proof.

Lemma 2. *Let $H(\infty, y) = y\,G(\infty, y) f(G(\infty, y))$ be the PGF for the distribution q_n, and suppose*

$$G(\infty, 1-w) \sim 1 - \phi(w) \quad \text{as } w \to 0, \tag{2.22}$$

where $w = 1 - y$. Then the asymptotic form of $H(\infty, y)$ is

$$H(\infty, 1-w) \sim 1 - w - (z+1)\phi(w) \quad \text{as } w \to 0, \tag{2.23}$$

and applying Lemma 1 allows the asymptotics of q_n to be determined from the nonanalytic part of $\phi(w)$.

If G asymptotically behaves as in Eq. (2.22), we can proof the lemma simply by expanding G in Taylor series to first order in ϕ noting that $\phi(0) = 0$ and $f'(1) = z$.

Using the asymptotic series 2.22, and combining the results of Lemmas 1 and 2, one can study the large-n asymptotic behaviour of q_n [28]. For the sake of clarity, we just present in Table 2.1 the final results of the asymptotic analysis of popularity distributions for some interesting particular cases. We observe power law dependence with different exponents for some of the analysed regimes.

Out-degree distribution	**Innovation rate**	**Asymptotic popularity distribution**
$p_k \propto k^{-\gamma}$, $2 < \gamma < 3$	$\mu = 0$	$q_n \sim n^{-\frac{\gamma}{\gamma-1}}$
$p_k \propto k^{-\gamma}$, $2 < \gamma < 3$	$0 < \mu < 1$	$q_n \sim n^{-\gamma}$
Finite $\sum_k k^2 p_k$	$\mu = 0$	$q_n \sim n^{-\frac{3}{2}}$
Finite $\sum_k k^2 p_k$	$0 < \mu < 1$	$q_n \sim n^{-\frac{3}{2}} e^{-\frac{n}{\kappa}}$, $\quad \kappa = 2\frac{2[2z+f''(1]}{\mu^2(z+1)^2}$

TABLE 2.1: Asymptotic popularity distribution for different out-degree distribution and innovation rates.

	p_k	μ	L
simulation (A)	$\delta_{k,10}$	0	10^5
simulation (B)	$k^{-2.5}$	0.01	10^5

TABLE 2.2: Simulation parameters. p_k is the out-degree probability distribution of the network; μ is the innovation rate, namely the probability of tweeting a new meme instead of retweeting the present content; L is the number of nodes (i.e. users).

2.3 NUMERICAL SIMULATIONS

For testing the goodness of the model and the validity of the old-age limit, numerical simulations are carried out for two out-degree distributions p_k, i.e. for two different network configurations. The system parameters considered in each simulation are summarized in Table 2.2.

The first simulation—from now on referred as Simulation (A)—is performed for $p_k = \delta_{k,10}$, for which we set $L = 10^5$ and $\mu = 0$, namely no innovation. Results are reported in Figure 2.7. Therein, we can observe that in the old age limit we get a good power law fit with exponent $\alpha = 0.51$ for the complementary cumulative distribution function (CCDF) $\sum_{n'=n}^{\infty} q_{n'}(a \to \infty)$. That means that $q_n\,(a \to \infty) \sim n^{-1.51}$, which shows an excellent agreement with the theoretical prediction in the case of finite second moment and vanishing innovation (see Table 2.1).

The second simulation—in the following, Simulation (B)—is instead performed for $p_k \propto k^{-\gamma}$ for $k > 4$ with $\gamma = 2.5$. For the aforementioned network, the second moment diverges in the large size limit. We have carried out the simulation in a system with $L = 10^5$ and $\mu = 0.01$. Again, in the old age limit, the CCDF is well described by a power law fit, in this case, with exponent $\alpha = 0.68$ for the CCDF. Therefore, we have obtain that $q_n\,(a \to \infty) \sim n^{-1.68}$. Remarkably, this does not agree with the theoretical prediction given by $q_n\,(a \to \infty) \sim n^{-2.5}$ (see Table 2.1). However, our numerical result approaches to the theoretical prediction for vanishing innovation $\mu = 0$ (see Table 2.1). We think the reason behind this disagreement may come from finite size effects. Note that the assymptotic behaviour is expected to be achieved for high popularities, and the cutoff due to the finite size of the network could prevent us to reach this asymptotic region. The finite size of the network also makes that the second moment is not formally infinity but simply very big.

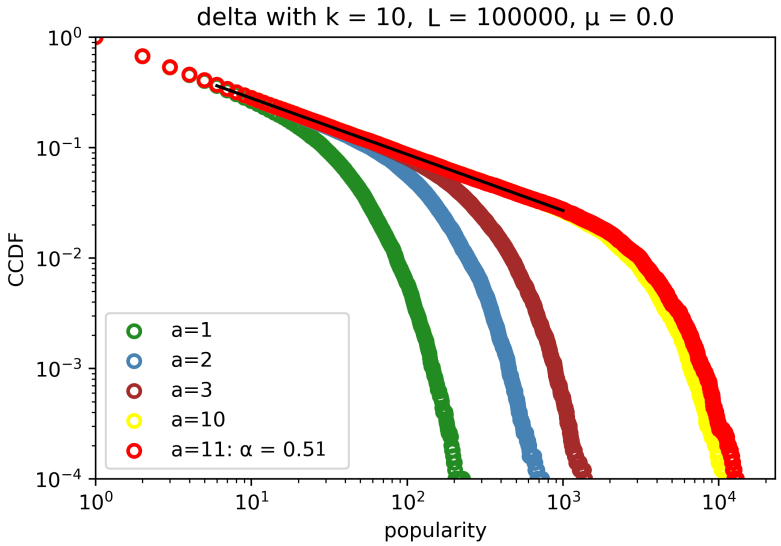


FIGURE 2.6: Numerical CCDF of the popularity for Simulation (A), where $p_k = \delta_{k,10}$ and $\mu = 0$. Numerical results are plotted using circles whereas the black line stands for the theoretical prediction.

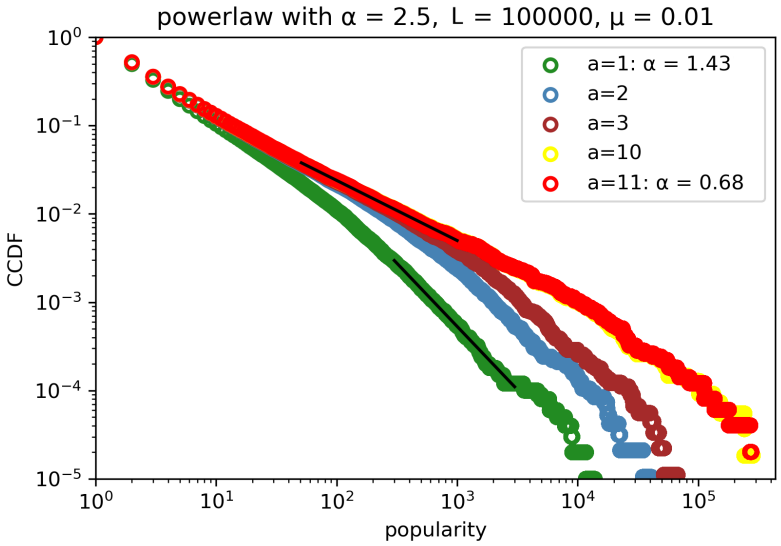


FIGURE 2.7: Numerical CCDF of the popularity for Simulation (B), where $p_k \propto k^{-\gamma}$ with $\gamma = 2.5$ and $\mu = 0.01$. Numerical results are plotted using circles whereas the black line stands for the theoretical prediction.

2.4 Meme Dynamics and Neutral Theory

So far, we have focused on the popularity of memes. At this point, we want to point out that the described model can be mapped into a stochastic neutral model for the meme's exposure, or equivalently in ecological terms, for the population dynamics of species.

Indeed, identifying memes as species and the number of screens in which they are present as the population's abundance, the time evolution of the number of different users following a given meme corresponds to studying the evolution of the abundances of species. Moreover, notes that our model is neutral since all the memes are equivalent and there is no preference on choosing a spreading meme, there is no fitter species. The only evolution driver is indeed demographic stochasticity, namely birth, death and speciation of memes in users' screens.

These are essentially the key ingredients of Neutral Theory, first formulated by Stephen Hubbel [31], that has achieve a remarkable success describing observed ecological patterns in very different ecosystems in spite of its simplicity which allows analytical treatment [53, 54, 55]. The main features of this theory are the following:

1. it assumes competitive equivalence among interacting species;
2. it is an individual-based stochastic theory founded on mechanistic assumptions about the processes controlling the origin and interaction of biological populations at the individual level ;
3. it can be formulated as a dispersal limited sampling theory;
4. it is able to describe several macro-ecological patterns through just a few of ecological processes as: birth, death and migration.

Therefore, its aim is not to describe the number of complex microscopic behaviours inside an ecological community, but rather at simplifying them by considering all the species as equivalent and their interactions as a stochastic process. In fact, the beauty of this theory relies in the attempt of capturing the fundamental aspects that characterise an ecosystem using a minimal number of ingredients. Although this can be a too simplistic and unrealistic assumption, as pointed out by many ecologists [19, 38, 40, 45], there are a lot of applications - from tropical forests [53] to coral reefs [55] - that show a strong agreement between Neutral Theory's predictions and observed ecosystems.

2.5 A Master Equation approach for Meme Exposure

In this section, we develop a neutral theory in order to derive the evolution equation for the distribution of meme exposure. By definition, the dynamics of the system is Markovian because, in an elementary time step, it depends only on the current state and not on the previous ones, that is, it has no memory. Furthermore, we are within the context of a neutral theory as discussed in Section 2.4 since all memes are equivalent, there are no fitter memes than others. Therefore, it is possible to derive the master equation that rules the memes' exposure evolution, i.e. that gives the probability density function $P(x,t)$ for a given meme to be present on a fraction x of the screens at time t.

Let us suppose we want to describe the dynamics of a given meme A among all. Then, we can identify the problem as a two-meme system where we define the state B as non-A. We recall that, at each time step, a random node is chosen and it can either spread its current meme - with probability $1-\mu$ - or create a new meme - with probability μ. Note that, since B is defined as non-A, if the random picked node innovates its meme, this will become B in any case. Therefore, if we denote by $k \in \{0, \ldots, L\}$ the number of the spreading node's followers and by $j \in \{0, \ldots, k\}$ the number of followers with a current meme different to the final meme of the spreading node, the transition rates that rule the dynamics are the following:

select and spread A this leads the system from x to $x + j/L$ and occurs with a transition probability

$$W_{k,j}^{A,s}(x) = x\,(1-\mu)\,p_k \binom{k}{j} x^{k-j}(1-x)^j$$

select and innovate A this leads the system from x to $x - (1+j)/L$ and occurs with a transition probability

$$W_{k,j}^{A,i}(x) = x\;\mu\;p_k \binom{k}{j} x^{j}(1-x)^{k-j}$$

select B this leads the system from x to $x - j/L$ and occurs with a transition probability

$$W_{k,j}^{B}(x) = (1-x)\,p_k \binom{k}{j} x^{j}(1-x)^{k-j}$$

To write these transition rates, one has to consider (i) the probability of selecting the spreading node either A (x) or B ($1-x$), (ii) then either innovate (μ) or not ($1-\mu$), (iii) then the probability of the selected node to have k followers (p_k), and (iv) finally the probability to take from this k followers the all possible combinations with memes A and B. In order to obtain a simpler expression for the transition rates, we have approximated the hypergeometric distribution, which corresponds to the real sampling without replacement, by the binomial distribution corresponding to sampling with replacement. This assumption is strictly valid for systems of large size with moderately low number of followers. The Master Equation then reads

$$\begin{aligned} P\,(x, t+\delta t) = \sum_{k=0}^{L}\sum_{j=0}^{k}\Big\{ & W_{k,j}^{A,s}\left(x - \frac{j}{L}\right) P\left(x - \frac{j}{L}, t\right) \\ & + W_{k,j}^{A,i}\left(x + \frac{1+j}{L}\right) P\left(x + \frac{1+j}{L}, t\right) \\ & + W_{k,j}^{B}\left(x + \frac{j}{L}\right) P\left(x + \frac{j}{L}, t\right) \\ & + \left[1 - W_{k,j}^{A,s}(x) - W_{k,j}^{A,i}(x) - W_{k,j}^{B}(x)\right] P\,(x,t)\Big\}. \end{aligned} \tag{2.24}$$

Assuming the innovation rate scales as $\mu = \tilde{\mu}/L$ and performing the typical Kramers-Moyal expansion (see Appendix B up to the second order, which is usually referred as the **diffusive approximation**, we get the Fokker-Planck Equation

$$\partial_\tau P(x,t) = -\partial_x[A(x)P(x,t)] + \frac{1}{2}\partial_x^2[B(x)P(x,t)] \tag{2.25}$$

where we have introduced a macroscopic timescale $\tau = t/L$ and neglected terms of order equal or higher to $1/L$. The drift and diffusion coefficients are

$$\begin{aligned} A(x) &= L\sum\left[jW_{k,j}^{A,s}(x) - (1+j)W_{k,j}^{A,i}(x) - jW_{k,j}^{B}(x)\right] \\ &= -\underbrace{\tilde{\mu}\left(1+\langle k\rangle\right)}_{a}x \equiv -ax, \end{aligned} \tag{2.26a}$$

$$\begin{aligned} B(x) &= \left[j^2W_{k,j}^{A,s}(x) + (1+j)^2W_{k,j}^{A,i}(x) + j^2W_{k,j}^{B}(x)\right] \\ &= \underbrace{\left[\left(1+\langle k\rangle\right)\langle k\rangle + \sigma_k^2\right]}_{b}x\left(1-x\right) \equiv bx\left(1-x\right), \end{aligned} \tag{2.26b}$$

where we have adapted to our case the general formula in Eq. (B.3).

Assuming the diffusive approximation to be valid, we therefore get a hard-to-solve Fokker-Planck Equation, due to the non linearity of the diffusion term $B(x)$. Indeed, exact full solutions for the Fokker-Planck equation are only known for very specific cases, like Ornstein-Uhlenbeck process.

2.5.1 AN IMPORTANT REMARK: DIFFUSIVE APPROXIMATION

At this stage, it is important to highlight that the diffusive approximation by which we have simplified Eq. (2.24) to Eq. (2.30) is not always valid. Indeed, a famous results by Kolmogorov and Lindeberg states the following

Theorem 1. *Let $X(t)$ be a continuous time Markov process whose range consists of the real numbers. For $t > 0$, let $P(x,t|\,x_0,0)$ be the probability density of $X(t) = x$ conditional on $X(0) = x_0$.*
As a consequence of the Markov property it obeys the Master Equation

$$\partial_\tau P(x,t|\,x_0,0) = \mathbf{W}\cdot P(x,t|\,x_0,0) \tag{2.27}$$

where W is a linear operator acting on the x-dependence. Its matrix element $\mathbf{W}(\mathbf{x}'|\mathbf{x})$ for $x' \neq x$ is the transition probability per unit time from x' to x and here is supposed to be time-independent.

Moreover, if the following three conditions holds

$$\lim_{\delta t \to 0} \frac{\langle X(t+\delta t) - X(t) \rangle}{\delta t} = A(x) \tag{2.28a}$$

$$\lim_{\delta t \to 0} \frac{\langle (X(t+\delta t) - X(t))^2 \rangle}{\delta t} = B(x) \tag{2.28b}$$

$$P\left(|X(t+\delta t) - X(t)| > \varepsilon\right) = O(\delta t) \quad \forall\, \varepsilon > 0 \tag{2.28c}$$

than the Master Equation is equivalent to the Fokker-Planck Equation

$$\partial_\tau P\left(x, t|\, x_0, 0\right) = -\partial_x [A(x) P\left(x, t|\, x_0, 0\right)] + \frac{1}{2} \partial_x^2 [B(x) P\left(x, t|\, x_0, 0\right)] \tag{2.29}$$

Although in many observed and well-studied systems the conditions (2.28) are easily satisfied, this is not true in general for the discussed model, since there is not a priori certainty that the last condition holds. Indeed, if we do not make any assumption on the out-degree, the selected spreading node may have so many followers that the exposure x of memes can change very sharply. These processes where abrupt changes occur are usually called bust processes and present a phenomenology different from that of diffusive processes. Therefore, if the dynamics is governed by such a bursts processes, we expect our diffusive approximation to fail.

2.5.2 MEAN FIRST PASSAGE TIME

In this section, we focus our attention on obtaining some exact results for our system. Let us consider our Fokker-Planck Equation (2.25)

$$\partial_\tau P\left(x, t\right) = -\partial_x [-axP\left(x, t\right)] + \frac{1}{2} \partial_x^2 [bx\left(1 - x\right) P\left(x, t\right)]. \tag{2.30}$$

We can rewrite it as a continuity equation

$$\partial_t P(x', t|x, 0) = -\partial_{x'} J(x', t|x, 0), \tag{2.31}$$

with the current given by

$$J(x', t|x, 0) = A(x') P(x', t|x, 0) - \frac{1}{2} B(x') P(x', t|x, 0). \tag{2.32}$$

This tells us how the probability density function evolves in time.

Herein, we focus on two quantities of great interest that can be studied. The first one is the probability $\pi(x)$ that a meme disappears before it reaches monodominance on the system, given that its initial exposure is x. The second one, instead, is the "Mean First Passage Time" (MFPT) $T(x)$ of the process. The time needed for the exposure to touch the boundary $x = 0$ before $x = 1$, conditioned to the fact of having an initial exposure x is a random variable, which has a specific probability distribution. The

MFPT is the first moment of that distribution. Note that, this observable contains insightful information of our system, allowing us to predict the timescale over the attention in a topic vanishes.

Following Gardiner [26], let us define some relevant probabilities. First, we define the probability of not exiting through a boundary at time t, given that we start from an initial exposure x,

$$g(x,t) = \int_0^1 dx P(x',t|x,0). \tag{2.33}$$

Note that $g(x,0) = 1$ since $P(x',0|x,0) = \delta(x'-x)$ and $\lim_{t\to\infty} g(x,t) = 0$ because if we wait enough the exposure always reach one of the boundaries. Studying the time evolution of this probability, we get

$$\int_t^\infty dt' \partial_{t'} \int_0^1 dx P(x',t'|x,0) = -\int_t^\infty dt' J(1,t'|x,0) + \int_t^\infty dt' J(0,t'|x_0,0). \tag{2.34}$$

Integrating over time, we obtain

$$\int_0^1 dx P(x,t|x_0,0) = \underbrace{\int_t^\infty dt' J(1,t'|x,0)}_{g_1(x,t)} \underbrace{- \int_t^\infty dt' J(0,t'|x,0)}_{+g_0(x,t)} \tag{2.35}$$

,where we have defined the probability of going through both boundaries $x = 0$ and $x = 1$ from time t, given that we start from x exposure, as $g_0(x,t)$ and $g_1(x,t)$ respectively. Note that $g_0(x,0) \equiv \pi(x)$ is one of the observables we are interested in, that is, the probability of exiting through $x = 0$, starting from a exposure x.

Since $P(0,t|x,0)$ satisfies a backward Fokker-Planck equation because the system is time translation invariant, as explained in Appendix C. Therefore, we get the evolution equation for $g_0(x,t)$

$$\begin{aligned} \partial_t g_0(x,t) = J(0,t|x,0) &= -\int_t^{+\infty} dt' \partial_{t'} J(0,t'|x,0) \\ &= A(x)\partial_x g_0(x,t) + \frac{1}{2} B(x) \partial_x^2 g_0(x,t). \end{aligned} \tag{2.36}$$

The probability that starting from exposure x, we reach absorption in $x = 0$, is a time $T > t$ reads

$$\text{Prob}(T > t) = \frac{g_0(x,t)}{\pi(x)} \tag{2.37}$$

Thus, the MPFT can be written as the integral

$$\begin{aligned} T(x) &= -\int_0^{+\infty} t \, d\text{Prob}(T > t) \\ &= -\int_0^{+\infty} t \, \partial_t \text{Prob}(T > t) dt \\ &= \int_0^{+\infty} \frac{g_0(x,t)}{g_0(x,0)} dt. \end{aligned} \tag{2.38}$$

Using Eq. (2.36) with respect to time t, we finally get a differential equation for the MFPT $T(x)$ coupled to $\pi(x)$,

$$A(x)\partial_x[\pi(x)T(x)] + \frac{1}{2}B(x)\partial_x^2[\pi(x)T(x)] = -\pi(x). \tag{2.39}$$

Thus, we first solve first $\pi(x)$ and afterwards we will return to (2.39).

By letting $t \to 0$ in Eq. (2.36), we notice that $J(0,0|x,0)$ must vanish if for any $x \neq 0$, since $P(0,0|x,0) = \delta(x)$. Hence, the right-hand side goes to zero and we get that $\pi(x)$ is the stationary solution of the backward Fokker-Planck equation, that is,

$$A(x)\partial_x\pi(x) + \frac{1}{2}B(x)\partial_x^2\pi(x) = 0. \tag{2.40}$$

$$\pi(x) = (1-x)^{1-\frac{2a}{b}}\,; \tag{2.41}$$

On the other hand, once $\pi(x)$ is solved, we can go back to (2.39), and taking into account the boundary conditions

$$\pi(x)T(0) = \pi(1)T(1) = 0, \tag{2.42}$$

we obtain for the MFPT

$$\begin{aligned} T(x) = \Bigg\{ & 2a\pi x(x-1)^{\frac{2a-b}{b}} \csc\left(\frac{2a\pi}{b}\right) - b\,{}_2F_1\left(1,1;1+\frac{2a}{b};\frac{1}{x}\right) \\ & +2ax\left[\gamma + \pi\cot\left(\frac{2a\pi}{b}\right) + \ln x + \psi_0\left(\frac{2a}{b}\right)\right]\Bigg\}\left[a(b-2a)\right]^{-1} \end{aligned} \tag{2.43}$$

where ${}_2F_1$, γ and ψ_0 stand for the generalised hypergeometric function ${}_pF_q$ with $p=2$ and $q=1$, the Euler's constant, and the digamma function respectively.

At the end of the day, we have a close expression for the MFPT in the diffusive approximation. The end of this section is devoted to the analysis of the dependence of the MFPT on the model parameters. As we can immediately notice from Eq. (2.43), the mean first passage time is a function of x and the parameters a and b given by Eq. (2.26). Note that we expect expression (2.43) to be valid for $2a < b$ since equation (2.41) verifies the correct boundary condition only in this case. It is handy to rewrite a and b as follows

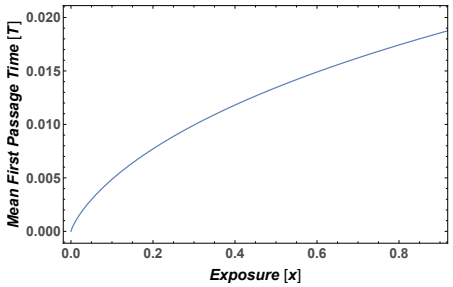
$$a = \tilde{\mu}\left(1 + \langle k\rangle\right), \tag{2.44a}$$

$$b = \langle k\rangle\left[(1+\langle k\rangle) + \langle k\rangle c_v^2\right], \tag{2.44b}$$

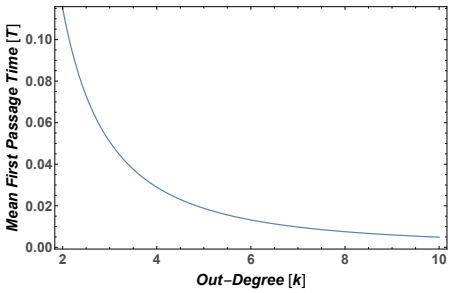
where we have introduced the usual definition of the coefficient of variation c_v. Therefore, we can study the MFPT as the function

$$T(x) = f\left(x;\langle k\rangle, c_v, \tilde{\mu}\right). \tag{2.45}$$

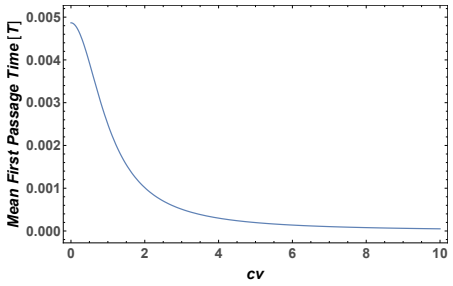
Fig.2.8 reports the dependences of T on its arguments. First, reasonably, we get a monotonically increasing function of x. The higher the initial exposure is, the higher



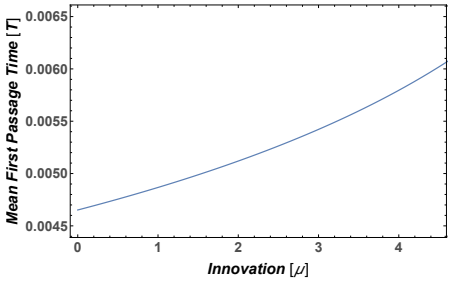
(A) Mean first passage time $T(x)$.



(B) Dependence of mean first passage time $T(x)$ on $\langle k \rangle$, namely $f\,(0.1; \langle k \rangle, 0, 1)$.



(C) Dependence of mean first passage time $T(x)$ on c_v, namely $f\,(0.1; 10, c_v, 1)$



(D) Dependence of mean first passage time $T(x)$ on $\tilde{\mu}$, namely $f\,(0.1; 10, 0, \tilde{\mu})$.

FIGURE 2.8: Mean first passage time $T(x)$.

the time to reach the exit through $x = 0$. The dependence of T on $\langle k \rangle$ is also intuitively explained. Higher $\langle k \rangle$ means a higher connectivity, which allows that invasions and removals of memes from the system occur faster. Analogously, if we keep $\langle k \rangle$ and we increase c_v, i.e. we increase σ_k, it is more likely that some nodes have an higher out-degree and the same reasoning applies. Nevertheless, the dependence on the innovation rate $\tilde{\mu}$ seems less intuitive for our understanding. Despite the fact that new memes are introduced in the system faster, we get that the time to loose the attention on a specific meme increases. It is true that it does not seem a very strong dependence for the explored values and also we have to take into account that the MFPT is not exactly the time to exit through $x = 0$ since it is conditionated to nor reaching monodominance first.

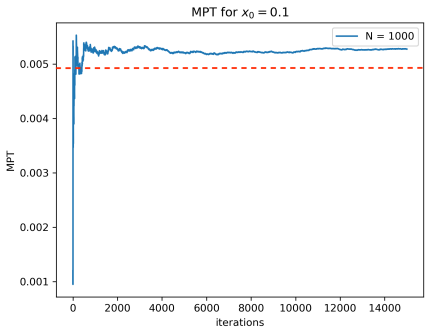
2.5.3 NUMERICAL SIMULATIONS FOR THE MFPT

We have performed simulations for checking the goodness of our results. First of all, we write the Langevin description corresponding to the Fokker-Planck equation (2.30)

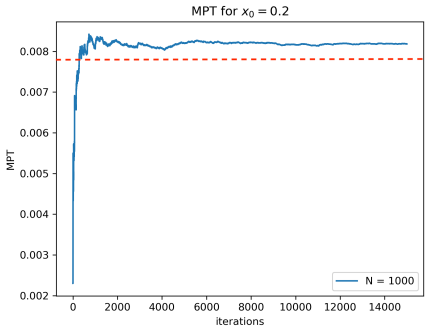
$$\dot{x}\,(t) = -ax + \sqrt{bx\,(1-x)}\,\eta(t), \tag{2.46}$$

where $\eta(t)$ is a Gaussian white noise with zero mean value and with time correlation $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. Langevin equation gives us the description of the stochastic trajectories of the process codified in a stochastic differential equation. Note that, in this

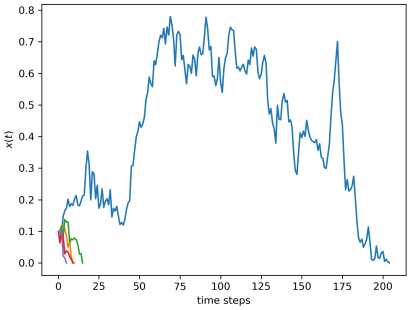
formulation, the drift term leads to a deterministic behaviour, whereas the diffusion term introduces the stochasticity at the the level of the dynamics of single trajectories.



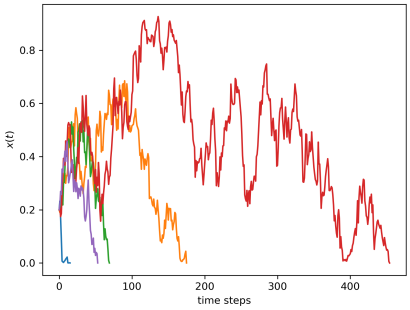
(A) Mean first passage time $T(x)$ versus number of trajectories simulated for computing it for $x = 0.1$. The dashed red line is the analytical prediction.



(B) Mean first passage time $T(x)$ versus number of trajectories simulated for computing it for $x = 0.2$. The dashed red line is the analytical prediction.



(C) Some trajectories for $x = 0.1$.



(D) Some trajectories for $x = 0.2$.

FIGURE 2.9: Simulations results for Langevin equations for $x = 0.1$ and $x = 0.2$.

Numerical integration of the Langevin equation (2.46) is reported in Fig. 2.9 and in Appendix D. As we can notice from both top panels, the number of trajectories needed for reaching a stable mean value for the passage time is around 5000. In the bottom panels, we show some specific trajectories. The numerical MFPT for different initial conditions x is compared to its theoretical prediction in Fig. 2.10, showing a perfect agreement with solution given by Eq. (2.43).

After having checked the necessary statistics for having a consistent mean, we performed simulations in our microscopic model for the two different configurations reported in Table 2.3. The first setting is $p_k = \delta_{k,10}$, $\tilde{\mu} = 1$ and $L = 1000$. Average results on over 5000 memes[3] are reported in Fig. 2.11. On the one hand, we observe that the results for small x, up to 0.2 are in perfect agreement with the theoretical predictions given by Eq. (2.43). Conversely, for x larger than 0.2 the is not quantitatively reliable

[3]More precisely, the number of simulations is such that at least 5000 memes are considered. Every point x along every trajectory $x(t)$ was used for computing the MFPT.

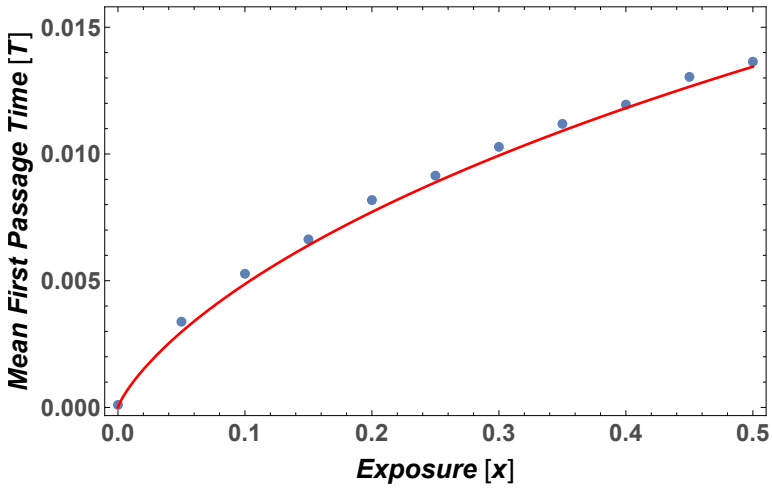


FIGURE 2.10: MFPT $T(x)$ for different initial conditions x for the process codified by the Langevin equation (2.46). Each point is a mean on 5000 trajectories.

	p_k	μ	L
simulation (C)	$\delta_{k,10}$	0.001	10^3
simulation (D)	$k^{-2.5}$	0.001	10^3

TABLE 2.3: Simulation parameters. p_k is the out-degree probability distribution of the network; μ is the innovation rate, namely the probability of tweeting a new meme instead of retweeting the present content; L is the number of nodes (i.e. users).

anymore and the observed mean first passage time is smaller than the predicted ones with our diffusive approach.

Furthermore, simulations with a different degree distribution, namely $p_k \propto k^{-2.5}$, but the same $\tilde{\mu} = 1$ and $L = 1000$, were also performed. Remarkably, this distribution has a mean value close 10, but the variance is much bigger than that in the previous case. Now, although we observe a reduction in the MFPT as it is also qualitatively predicted by our theory, we obtain no quantitative agreement for the whole range of initial exposures x, as shown in Fig. 2.11.

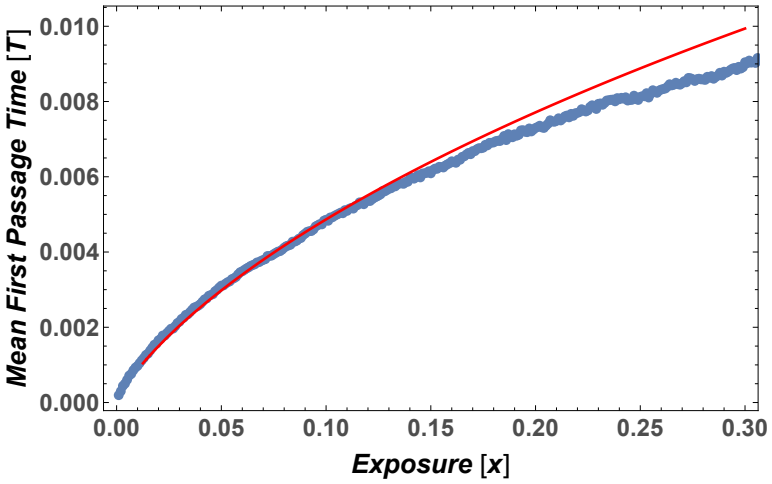


FIGURE 2.11: MFPT $T(x)$ for different x for simulation (C), namely for $p_k = \delta_{k,10}$ and $\tilde{\mu} = 1$. Each point is an average.

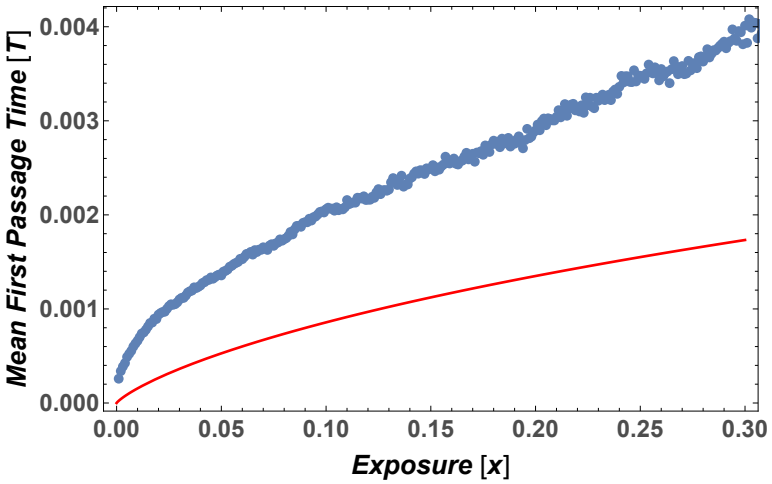


FIGURE 2.12: MFPT $T(x)$ for different x for simulation (D), namely for $p_k \propto k^{-2.5}$ and $\tilde{\mu} = 1$. Each point is an average.

2.5.4 APPROXIMATE FOKKER-PLANCK EQUATION AND MEAN EXPOSURE

As already pointed out in Section 2.5, even if some results as the mean first passage time can be obtained, the quadratic term x^2 in B in the Fokker-Planck equation (2.30) makes the equation very hard to solve. However, since we are modelling meme exposure and it is realistic that a single meme appears in a small fraction of the screens, we consider in this subsection that the quadratic term is negligible compared to the linear one, that is,

$$B(x) = bx + O(x^2). \tag{2.47}$$

Therefore, assuming that the exposure $x(t)$ of a meme varies smoothly with time and is not too large we get the approximated Langevin equation

$$\dot{x}\,(t) = -ax + \sqrt{bx}\,\eta(t). \tag{2.48}$$

Coming back to the Fokker-Planck description, we get

$$\partial_\tau P\,(x,t) = a\,\partial_x[P\,(x,t)] + \frac{1}{2}b\,\partial_x^2[xP\,(x,t)]. \tag{2.49}$$

We considere here absorbing boundary $x = 0$ since once a meme disappears from the system, i.e. x touches 0, there is no mechanism of re-introduce it. Remarkably, this Fokker-Planck equation is the same one that Azaele *et al.* [6] have found in a neutral ecological model.

Therein, they solve exactly the equation and get the time dependent solution for initial exposure x_0

$$P\,(x,t|\,x_0,0) = \frac{2a}{b}\frac{1}{1-e^{-at}}\exp\left[-\frac{2a\left(x+x_0e^{-at}\right)}{b\,(1-e^{-at})}\right]\sqrt{\frac{x_0}{x}e^{-at}}\,I_1\left[\frac{4a\sqrt{x_0xe^{-at}}}{b\,(1-e^{-at})}\right]. \tag{2.50}$$

Solution (2.50) provides us with the full statistical information of the our approximated Fokker-Planck. Specifically, we can compute all the moments from it. Herein, we focus here on the first moment, that is, the mean exposure. It is possible explicitly compute $\langle x\rangle_{x_0}$ by integrating x times $P\,(x,t|\,x_0,0)$, obtaining

$$\langle x\rangle_{x_0} = x_0\exp\left[-at\right]. \tag{2.51}$$

This is exactly the same result that could be obtained by considering only the deterministic part of the Langevin equation. Although this solution in which the noise term does not play any role may seem too simplistic, numerical simulations clearly show a good agreement with the predicted $\langle x\rangle_{x_0}$, specially for small ages.

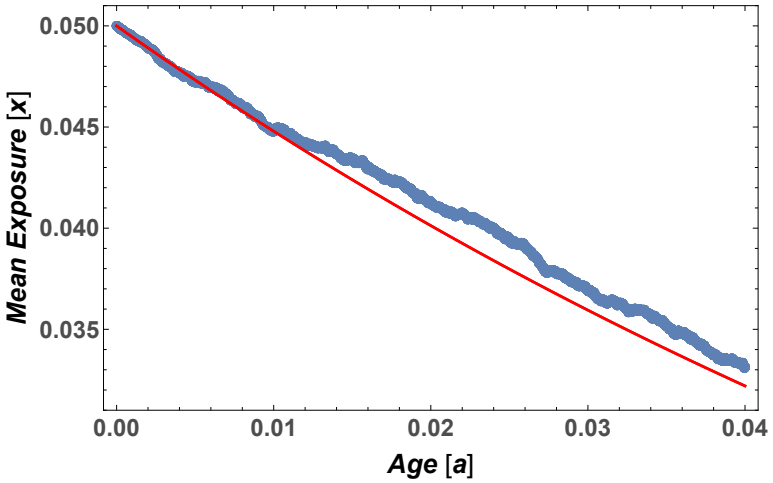


FIGURE 2.13: $\langle x \rangle_{x_0}$ for simulation (C), namely for $p_k = \delta_{k,10}$ and $\tilde{\mu} = 1$. The average is computed among all the memes that did not reach the absorbing state $x = 1$. Numerical results are plotted using blue circles whereas the red line stands for the theoretical prediction.

CHAPTER 3

POPULARITY ON YOUTUBE

YouTube represents one of the largest scale and most sophisticated recommendation systems in existence [20]. Indeed, its machine learning algorithm is responsible for recommending personalised content from an ever-growing corpus of videos to more than a billion users.

One of the most interesting properties of Youtube videos is the number of views, i.e. the times a video has been watched by different users. This can be informative for estimating if a video is, or will go, "viral", where with virality here we mean the probability for a video to be watched [29].

Videos with similar content may have a very different amount of views. How can this happen? Why does a video become popular, while some others not? Which features should a video have for being watched by millions of people? Can we predict popularity?

Answering these questions would not be so challenging if we knew Youtube's algorithm. Unfortunately, YouTube's recommendation system draws from machine learning techniques that are proprietary. To study how the virality arises, and to partially overcome this problem, we simulated by mean of a crawler - described in next Section - the behaviour of a user interested in the 2018 Italian General Election [1]. The crawler explores political videos following YouTube recommendations few weeks before the election day. The result of this search is a set of data containing videos appearing on YouTube collected while crawling.

In this chapter, we analyse these data aiming at understanding YouTube's recommendation system. In particular, we want to find out if there is any correlation between the structural properties of the network built using these data and the virality of videos.

3.1 YOUTUBE CRAWLER

A Web crawler, sometimes also called a spiderbot, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing [43]. A Web crawler developed by G. Chaslot [36]—which from now on will be referred as "YouTube Crawler"—was used for simulating the behaviour of a YouTube user and gathering recommendation data on the five main candidates of the Italian General Election in 2018: Silvio Berlusconi, Luigi Di Maio, Pietro Grasso, Matteo Renzi and Matteo Salvini. During the three weeks before the election, it acted as a YouTube user interested in one of the main candidates. It discovered a video through a YouTube search, and then followed a chain of YouTube recommended titles appearing *up next*, that is, within the first items of the recommendation list.

More specifically, the crawler ran 18 days over a time span of 21 days (three days of data are missing) during the 3 weeks before the election as follows.

1. The software searched for each candidate separately (each candidate corresponds to a different sample). It then clicked on several search results (the first 4 videos in the recommendation list) and captured which videos YouTube was recommending up next.
2. The process was then repeated, this time by selecting a sample of those videos YouTube had just placed up next. The sample was collected among the 19 videos again selecting the top 4 (*branching* of the search), with the constraint of not choosing any video twice.
3. Step 2 was repeated going deeper and deeper on the recommendations up to 4 times (*depth* of the search).

Hence, in a single-user trajectory, the crawler visits a maximum of 5 videos. This is one of the possible $4^4 = 256$ trajectories generating from a search keyword using a *branching* of 4 for a *depth* of 4. Each one of these is simulated by the crawler and during these explorations data are collected.

By design, the program operated without using a user's viewing history, ensuring it was capturing exclusively generic YouTube recommendations rather than those personalised to individual users. For every sampled video, the program collected the number of views, likes, dislikes, and recommendations, i.e. the number of times that the video was recommended during the crawling. Furthermore, the algorithm saved for each video a list with the recommendations appearing on screen for that video. Some examples of the data are reported in Table 3.2, while sample sizes are reported in Table 3.1.

We have collected the number of views and likes for every crawled video. This allows to study statistics about the explored data. In particular, the distribution of number of views and likes were collected for everyday and every initial keyword. As an example, data for initial keyword "Matteo Renzi" for the first day are reported in Fig. 3.1, the rest of them being reported in Appendix E. As we can see, number of views, as well as likes, are broadlydistributed, ranging from a few to millions. This means that YouTube does not recommend only the most viewed or top-rated videos. Furthermore, as the crawler ran for almost three weeks, some videos were recommended for more than just one day. This allows to study the *dynamics* of the number of views in time.

day	Berlusconi	Di Maio	Grasso	Renzi	Salvini	total
0	923 (228)	992 (244)	1014 (239)	890 (221)	705 (167)	1864 (780)
1	994 (241)	572 (93)	969 (242)	755 (149)	688 (125)	1737 (635)
2	895 (209)	801 (176)	1006 (251)	764 (184)	829 (206)	1924 (778)
3	973 (220)	967 (235)	1037 (255)	976 (235)	917 (215)	2086 (869)
4	958 (225)	859 (211)	986 (241)	972 (196)	607 (111)	1773 (667)
5	1031 (231)	1043 (240)	1108 (244)	1086 (229)	773 (145)	2001 (765)
6	881 (207)	696 (143)	936 (232)	945 (201)	0 (0)	1667 (561)
7	1036 (228)	1005 (241)	1092 (254)	1017 (219)	326 (54)	1907 (702)
10	1134 (248)	1003 (235)	1060 (236)	997 (213)	717 (125)	1995 (756)
11	1116 (241)	972 (222)	1138 (257)	986 (198)	708 (126)	1953 (710)
12	1076 (255)	959 (207)	1141 (243)	1083 (216)	614 (106)	2037 (721)
13	1099 (249)	974 (212)	1085 (234)	1017 (213)	822 (156)	2030 (747)
14	1077 (241)	975 (190)	951 (194)	1018 (212)	742 (142)	1903 (664)
16	1037 (225)	826 (159)	1032 (212)	1000 (211)	731 (153)	1848 (636)
17	1080 (241)	944 (186)	1092 (234)	990 (206)	882 (175)	1906 (674)
18	1051 (217)	1059 (232)	1032 (229)	925 (184)	976 (181)	1947 (677)
19	1024 (224)	1008 (215)	1018 (219)	900 (196)	939 (215)	1933 (725)
20	1037 (241)	997 (187)	1033 (222)	891 (178)	721 (118)	1819 (614)

TABLE 3.1: Number of videos reported by the crawler in each day and politician's initial search keyword. In parenthesis the number of video effectively explored (the maximum number is $4^4 = 256$), i.e. for which we have collected the number of views and have out-degree greater than zero. The last column corresponds to the total number of explored videos for that particular day: this is lower or equal than the sum of the other columns, since the same video could be explored by starting from different initial keywords. Data for "Matteo Salvini" on day 6 are missing.

id video	title	channel	recommendations	views	likes	dislikes
Wyi9QVOL2G8	Matteo Renzi a Otto e Mezzo	Matteo Renzi	5dzgIAbE9g, rm5t6X2YvNk, ..., OsAY-5T7tkQ	930	61	34
5dzgIAbE9g	Otto e mezzo 09 02 2018	La7	Wyi9QVOL2G8, FrO28akOmzA, gp4m5mrsmjg, mp9Ppep9eU, ... , XDfmhgFQ1CQ	1240	108	18
WqF0Fayjd6Y	Luigi Di Maio (M5S) a Canale5	piccolojason	C8jYhDMk06c, 91sc5q3HLIM, ..., 9WhWFhBJztE	19938	370	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮
mvCd9zoczhc	Renzi a Porta a porta su programma PD	michele vacca	wck1RVyvPM, E0aPiMaZCrY, ..., wvhKUQYvTP4	578	9	2

TABLE 3.2: Some of the 221 videos collected for Matteo Renzi on first day.

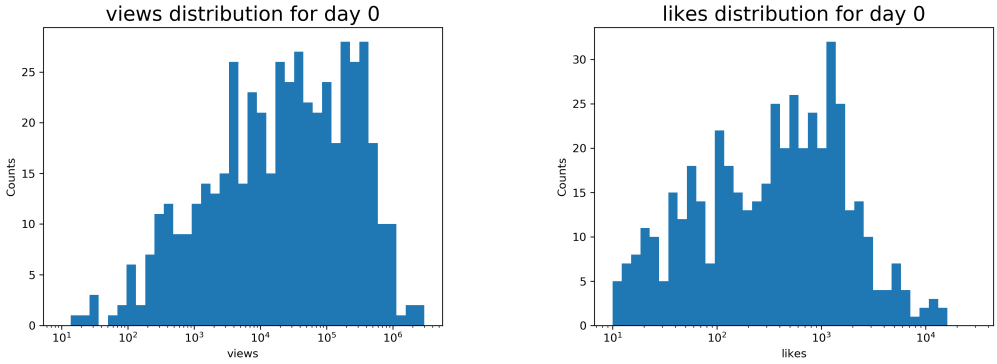
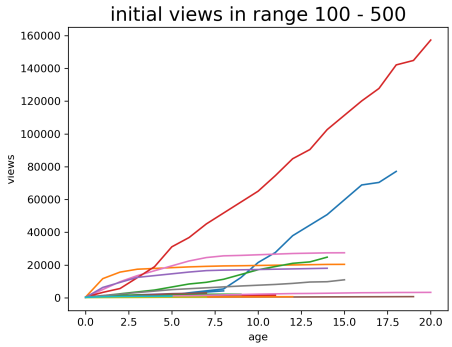


FIGURE 3.1: Views and Likes distribution of videos collected on day 0 for "Matteo Renzi" as initial keyword.

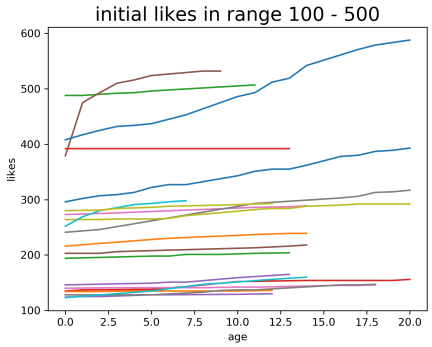
Fig. 3.6 reports the evolution of the number of views and likes for some randomly selected videos, classified by their initial visualisations, i.e. views in the first day in which the crawler found them. As we can see, different behaviours can be observed. Many videos remain almost stable, or increase the number of their views just by a little; whereas others increase exponentially in time as they become popular. This latter behaviour is usually related with *virality*.

It is natural to consider each collected sample as a network: videos are nodes connected by directed links representing recommendations from a node to another. Hence, a natural framework for studying these datasets is Network Theory, introduced in Chapter 1.3. For instance, we have studied both in-degree and out-degree distributions for Matteo Renzi in the first day in Fig. 3.3 and Fig. 3.4, respectively.

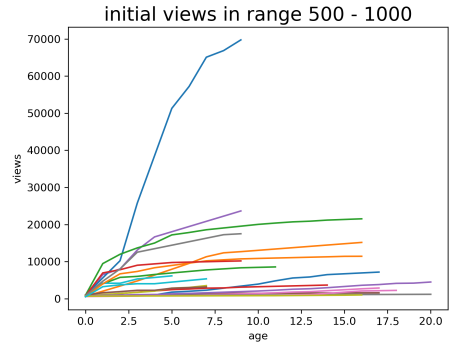
The insets show a snapshot of the corresponding graphs for all the videos explored by the crawler. In both graphs we report only nodes with non-zero out-degree, i.e. nodes that have been visited by the crawler. Similar plots for the rest of days and politicians are reported in Appendix E. Note that nodes in the insets are coloured according to labels. Specifically, videos were labelled depending on their title or their channel: if the name of a politician, his party or any of other politician related to him appears in the title, a tag (or label) was added to the node. If no name appears in a video, no tag was added to it. If more than one name appears in a video's title, we added multiple tags.



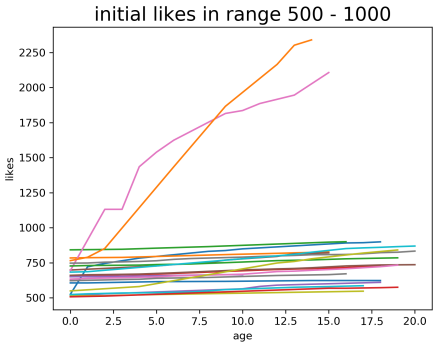
(A) Views versus age for some of the videos with initial views between 100 and 500.



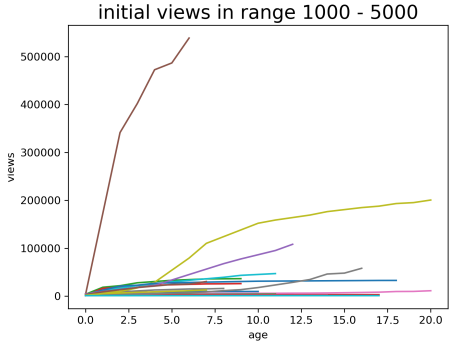
(B) Likes versus age for some of the videos with initial views between 100 and 500.



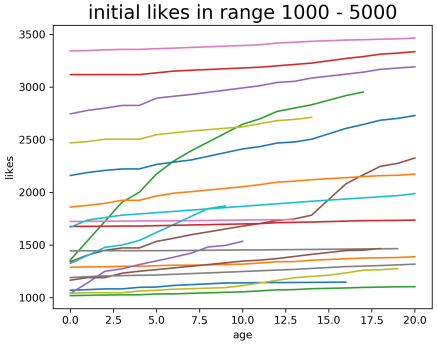
(C) Views versus age for some of the videos with initial views between 500 and 1000.



(D) Likes versus age for some of the videos with initial views between 500 and 1000.



(E) Views versus age for some of the videos with initial views between 1000 and 5000.



(F) Likes versus age for some of the videos with initial views between 1000 and 5000.

FIGURE 3.2: Views and Likes versus age for some of the collected videos. The age of a video is the defined as the time (in units of days) after its first appearance in the crawler.

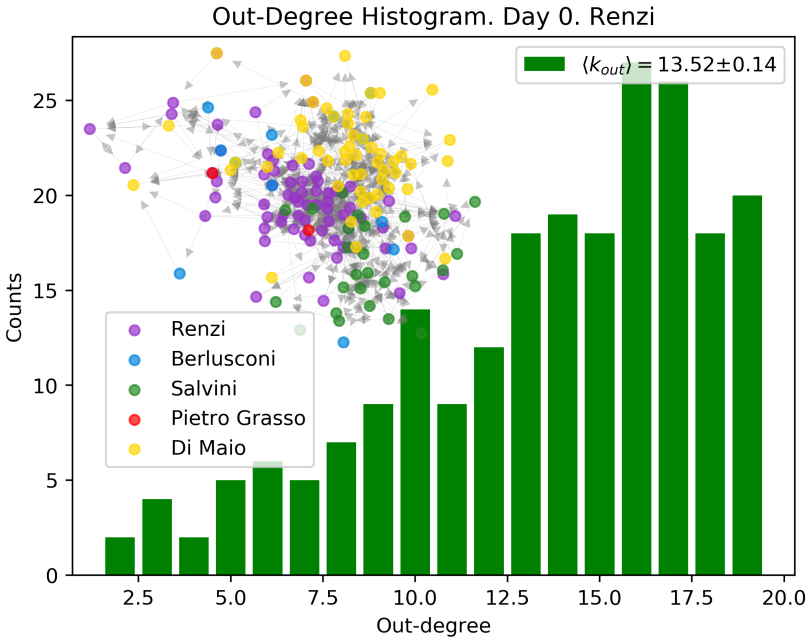


FIGURE 3.3: Out-degree distribution of videos collected on day 0 for "Matteo Renzi" as initial keyword.

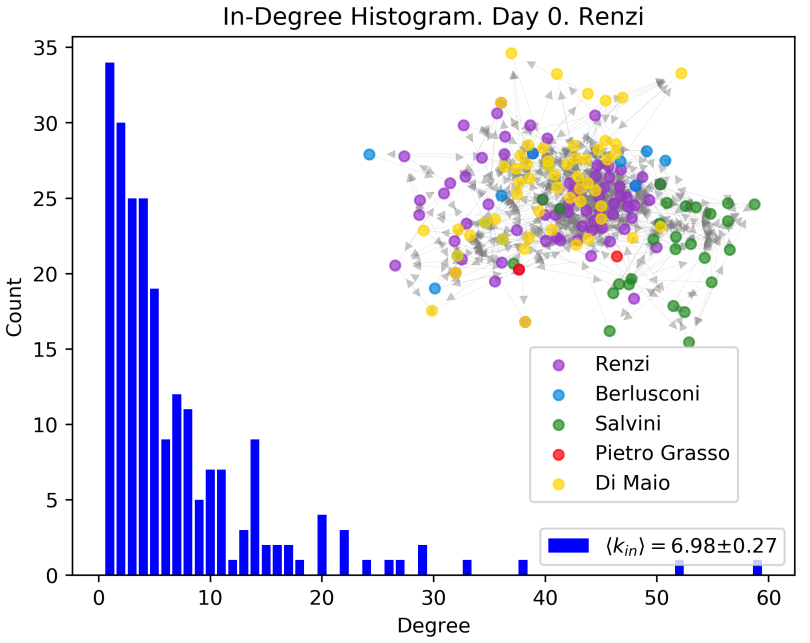


FIGURE 3.4: In-degree distribution of videos collected on day 0 for "Matteo Renzi" as initial keyword.

3.2 AVERAGE NETWORK

As seen in Fig. 3.4, and also in the similar plots reported in Appendix E, there are very few nodes with a high in-degree, the majority having a small in-degree. The formers are more recommended. It is thus natural to think that the most recommended videos are more likely to have a larger number of views. Intuitively, if these networks were stable, for a user who follows recommendations, the probability of viewing these videos would have been higher than for the others. Therefore, we expect some correlation between nodes' centrality and the number of views.

In this spirit, we have investigated possible correlations between the intrinsic structural properties of the reconstructed graphs and the number of views. In particular, our hypothesis is that each data sample collected for a particular initial keyword could be interpreted as a subsample of the total YouTube recommendations graph for that particular day. Therefore, for each day we have built the corresponding graph. This is the union of the five graphs corresponding to the candidates names. Then, we examine the correlation between different definitions of centrality and views for each particular day.

We start our analysis by the in-degree centrality. As already defined in 1.3, the in-degree centrality of a node is the number of nodes its incoming edges are connected to, normalised by the total number of nodes. Results for the first day can be observed in Fig. 3.6a. The corresponding results for the rest of the days are reported in Appendix E. Apparently, and against our intuition, no clear pattern of correlations is evident from these plots. Moreover, one could think that a highly recommended video is more likely to become viral. However, looking at time evolution of number of views for the most central videos, as can be seen in Fig. 3.5, none of them becomes viral.

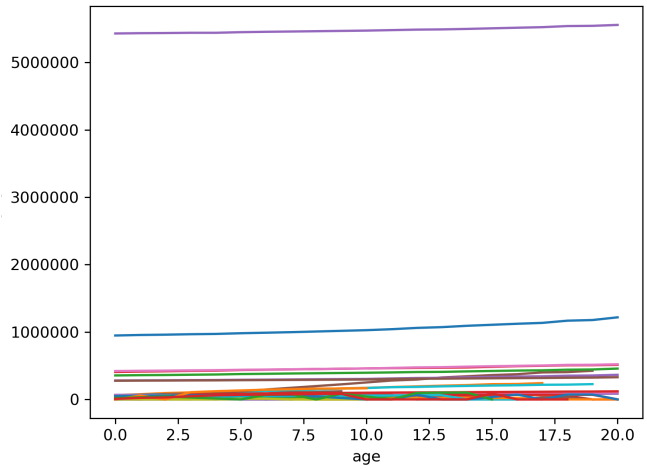
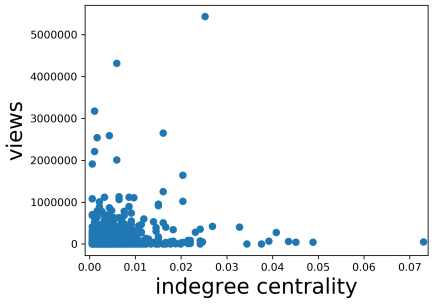


FIGURE 3.5: Time evolution of number of views for all the videos with an in-degree centrality greater than 0.03.

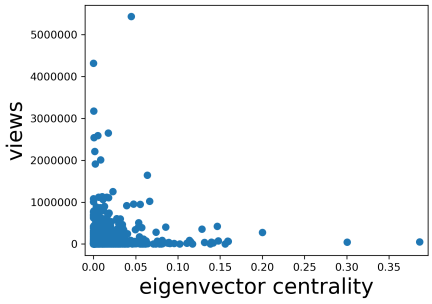
One of the possible explanations for this lack of correlation is that the in-degree

centrality is not the most appropriate measure of centrality. For instance, it could happen that although a video has a high in-degree, it is recommended only by low-in-degree videos. In that sense, a more suitable definition of centrality is the so-called "eigenvector centrality". This computes the centrality for a node based on the centrality of its neighbors. In this case, the relevance of a node is increased by being connected to other nodes that are themselves important.

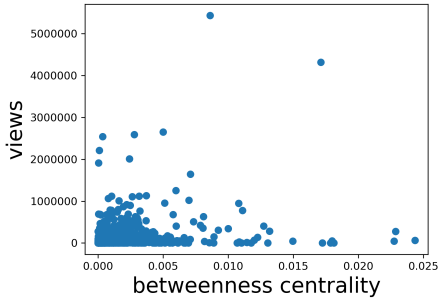
There are of course many alternative definitions of importance, and correspondingly many other centrality measures for networks. Two of the most used are "Betweenness centrality" [41] and "PageRank centrality" [16]. The former is related to the number of shortest paths between any two nodes passing through a given focal node. The latter is a variant of the eigenvector centrality that dilutes the centrality contribution of neighbours of a given node by normalising them by their out-degree.



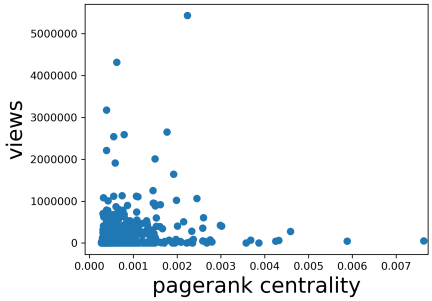
(A) Number of views versus indegree centrality for videos collected on first day.



(B) Number of views versus eigenvecetor centrality for videos collected on first day.



(C) Number of views versus betweenness centrality for videos collected on first day.



(D) Number of views versus PageRank centrality for videos collected on first day.

FIGURE 3.6: Number of views versus different centrality.

In spite of the various definitions of centrality used in our analysis, we have found no correlation between any of these measures and the number of views. This is shown in Figures 3.6b, 3.6c and 3.6d.

3.3 TEMPORAL NETWORK

So far, we have considered separately each average network for each day, without considering any relationship between them. Indeed, all the previously discussed measures of centrality are static and refer only to a particular day: each day, a node is ranked independently from its previous history. Nevertheless, we expect that videos that are recommended for several days are more likely to become viral than videos that appear only one day; even if they are more central on that particular day. Furthermore, YouTube recommendation system evolves in time, suggesting different videos from day to day, even if we start from the same video. In the language of networks, this means that YouTube is a temporal network: links can appear and disappear in time. These two considerations lead us to consider a *dynamical* model. The goal is to capture how past history affects quantities as the number of views.

For this purpose, we considered a model recently proposed by Della Vecchia et al. [23], "Dynamic SpringRank". This model is an extension of the standard (static) "SpringRank" algorithm by De Bacco et al. [21] that incorporates time information. In its static version, this physically-inspired model aims at ranking hierarchically nodes in directed networks.

Given the adjacency matrix A, our goal is to find a ranking of the nodes. SpringRank model computes the optimal location of nodes in a hierarchy by considering the network as a physical system. Specifically, each node i is mapped to a real-valued position or rank s_i (a *latent* variable), and each directed edge $i \to j$ becomes an oriented spring with a nonzero resting length and displacement $s_i - s_j$. Since we are free to rescale the latent space and the energy scale, we assume that both the spring constant and the resting length are set to 1. Thus, the spring corresponding to an edge $i \to j$ has energy:

$$H_{ij}(s_i, s_j) = \frac{1}{2}\left(s_i - s_j - 1\right)^2 \tag{3.1}$$

which is minimized when $s_i - s_j = 1$. According to this model, the optimal rankings of the nodes are the ranks $s^* = (s_1^*, \dots, s_N^*)$ which minimize the total energy of the system given by the Hamiltonian:

$$H(s) = \sum_{i,j=1}^{N} A_{ij} H_{ij}(s_i, s_j) = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij} \left(s_i - s_j - 1\right)^2 . \tag{3.2}$$

This model has been generalised [23] to account for time-varying ranks s_i^t. We use this recent development and infer dynamical-SpringRank ranks s_i^t for each node in the dynamical networks. These are proxy for a dynamical notion of centrality in a network.

Correlation between final numbers of views and scores are reported in Fig. 3.7. As we can see in Fig. 3.7, we found no clear correlation pattern between this centrality measure and the number of views again. We do not observe any correlation even if we focus just on viral videos, as depicted in Fig. 3.9. All these observations clearly show that YouTube is far from being a stable network, i.e. suggesting always the same videos and ranking them just by their number of views.

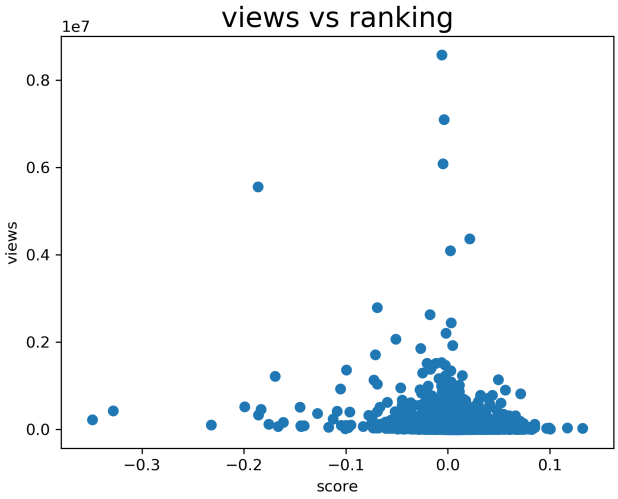


FIGURE 3.7: SprinRank scores versus final number of views.

3.4 ASSORTATIVITY ANALYSIS

Although there is no evidence that YouTube recommended videos are more likely to become popular, it is insightful to investigate the relation between recommendation and popularity patterns. As already mentioned, tags corresponding to the politicians were added to the nodes, according to the name and the channel of every investigated video. Interestingly, even if the crawler started its exploration on YouTube with an initial research of one of the political leaders, not all the suggested video are tagged according to the searched name. Looking at Fig. 3.4, whose inside report the videos' network corresponding to data for initial keyword "Matteo Renzi" for the first day, we notice immediately that also videos with tags different from Renzi are suggested.

Furthermore, not only videos related to different politicians are suggested regardless the initial searched name, but graphs seem to show a tendency on suggesting videos that share the same tags. Looking at Table 3.3, in which every entry (x, y) is the fraction of recommendation from videos with tag x to videos with tag y, we clearly see that the elements on the diagonal, corresponding to suggestions between videos related to the same politician, are at least one order of magnitude larger than the off-diagonal elements.

As defined in section 1.3.3, this tendency can be quantified by mean of the assortativity r. We recall that $r = 1$ for a completely assortative graph, $r = -1$ for a completely disassortative graph and $r = 0$ for a random graph [41]. Assortative measures are reported in Table 3.4, showing a strong assortativity.

Surprisingly, the assortativity is almost the same for every sample, regardless the initial searched name, and also for the average networks. Moreover, for the latter ones, assortativity is upheld by observing that if we either shuffle the tags of the nodes or the edges by holding the out-degree constant r becomes almost zero, as we expect for a random graph.

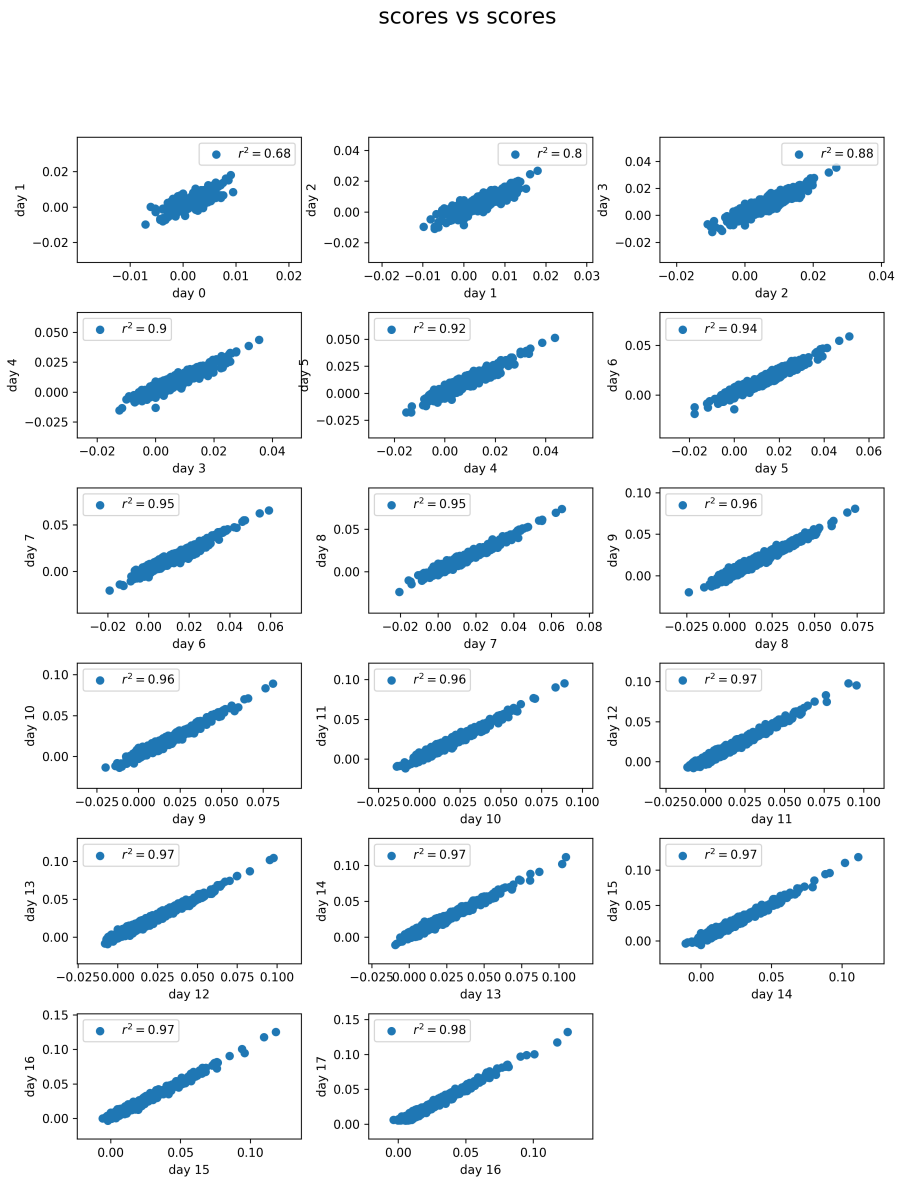


FIGURE 3.8: Correlation between scores in the "Dynamic SpringRank" model for consecutive days.

	Berlusconi	Di Maio	Grasso	Renzi	Salvini
Berlusconi	0.1443	0.0127	0.0005	0.011	0.0076
Di Maio	0.0145	0.313	0.0005	0.0181	0.011
Grasso	0.0037	0.0034	0.038	0.0083	0.0024
Renzi	0.012	0.023	0.0047	0.1643	0.0157
Salvini	0.0069	0.0113	0.0007	0.0096	0.1629

TABLE 3.3: Mixing between different candidates' videos for video collected on the first day. Only videos with one tag are considered.

day	Berlusconi	Di Maio	Grasso	Renzi	Salvini	Unfied
0	0.74	0.6	0.66	0.73	0.72	0.76
1	0.79	0.61	0.68	0.74	0.76	0.77
2	0.76	0.66	0.71	0.69	0.79	0.78
3	0.78	0.71	0.66	0.78	0.8	0.78
4	0.7	0.73	0.64	0.72	0.71	0.74
5	0.72	0.73	0.65	0.7	0.71	0.76
6	0.7	0.66	0.57	0.64		0.71
7	0.73	0.77	0.65	0.68	0.69	0.76
10	0.78	0.73	0.66	0.75	0.7	0.77
11	0.77	0.7	0.63	0.75	0.67	0.74
12	0.75	0.72	0.65	0.75	0.7	0.77
13	0.79	0.75	0.63	0.78	0.73	0.77
14	0.77	0.7	0.67	0.76	0.75	0.78
16	0.72	0.7	0.69	0.71	0.68	0.75
17	0.71	0.6	0.7	0.7	0.68	0.73
18	0.77	0.69	0.63	0.76	0.75	0.76
19	0.75	0.67	0.66	0.75	0.69	0.76
20	0.73	0.68	0.69	0.75	0.73	0.76

TABLE 3.4: Assortativity for graphs corresponding to videos collected for different candidates as initial keyword on different days. Last column shows the assortativity for the unified graph.

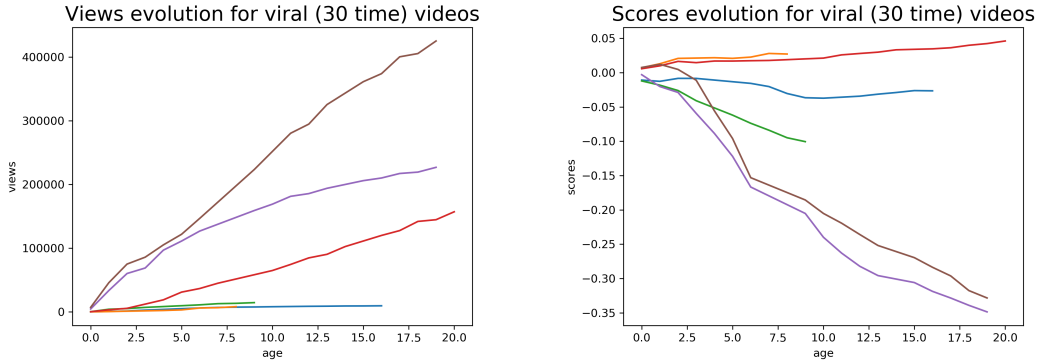


FIGURE 3.9: Views and scores evolution for viral videos, here defined as videos that incremented their number of views from the first to the final appearance by a factor 30.

These results show that starting from a given node with a given tag (representing a political party), YouTube recommendation system will most probably ($r = 0.75$!) suggest a video of the same tag, i.e. supporting the same political party. In this way we have a mechanisms for which the user tends to see only videos that are similar among them, i.e. reinforcing and reflecting similar content of the first video.

Therefore the YouTube recommandation systems can be considered an "Echo chamber". Although we lack of a quantitative and rigorous definition of echo chamber, we can consider it as mechanisms in OSN where users are somehow induced to prefer to interact only with ideologically-aligned peers. Although positive assortative mixing (also known as homophily) is one of the characteristic features of OSN [34] and reflect the natural phenomena where individuals prefer to interact with other individuals who are similar to them, echo-chambers exacerbate this phenomena and induce even stronger polarisation of societies into groups with different perspective [3]. Moreover it is also believed that echo-chambers tend to facilitate misinformation spreading and contribute to radicalise political discourse [22].

In this framework, a way to measure the influence of a politician could be looking at how much his videos "break" the echo chambers, namely what is the probability that one video related to him is suggested from videos related to one of the others. A possible measure of this influence is the following quantity

$$f_y = \frac{\sum_{x \neq y} e_{xy}}{\sum_x e_{xy}} = \frac{\sum_{x \neq y} e_{xy}}{b_y} \tag{3.3}$$

which is the fraction of videos recommending videos related to a political leader y, that are not related to y, normalised by the total fraction of videos pointing to y. Such a quantity, here defined as **mixing**, is represented in Fig. 3.10. As we can see, f is larger for Matteo Renzi and Luigi Di Maio, showing that these two politicians are more recommended than the others.

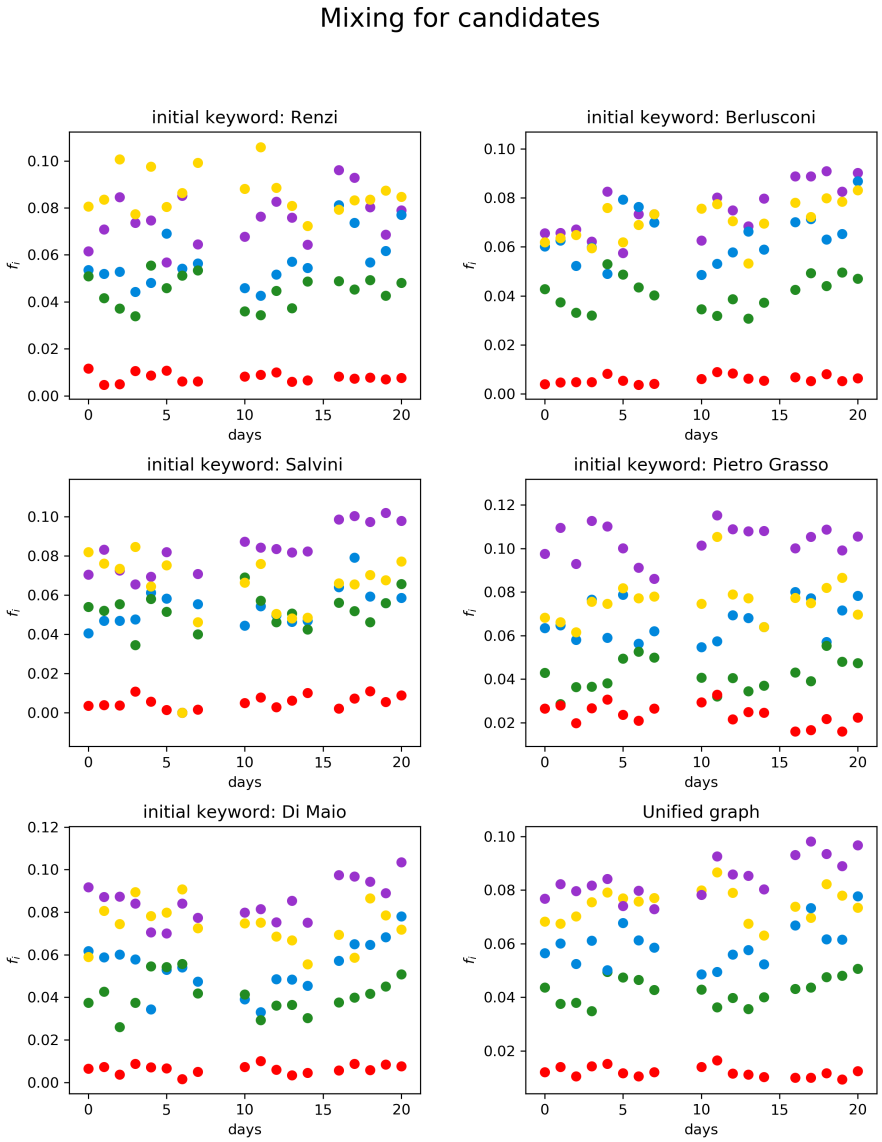


FIGURE 3.10: Mixing for every politician on every sampled day for the relative graph. The colour notation is the same as the one of previous figures: blue for Berlusconi, yellow for Di Maio, red for Grasso, purple for Renzi and green for Salvini.

CHAPTER 4

CONCLUSIONS

In this thesis we have proposed a theoretical framework rooted in statistical mechanics and network theory to study and analyze emergent patterns and social interactions in online social networks. Specifically, we have followed an ecological approach to analyse an individual based model for Twitter dynamics, whereas we have studied real data from YouTube through the lense of network theory.

In chapter , we have put forward a minimal model inspired by Twitter for the spreading of memes. This was originally introduced by Gleeson et al. [28], who focus on the study of popularity. In Section , we have reviewed Gleeson's results, showing that in some limits, popularity distribution asymptotically follows a power-law distribution, the exponent of which is related with the network topology. Numerical simulations for different networks configurations are reported in section 2.3, where we find a significant agreement with theoretical predictions.

Then, we have focused our attention on the *exposure*, which is a measure of the level of attention to a given meme. We have developed a master equation approach strongly inspired in the neutral theory for ecosystems. From the master equation, we have performed a diffusive approximation that leads to a Fokker-Planck equation for the probability density function of the exposure of a given meme.

Finding a full solution for the aforementioned Fokker-Planck equation entails a mathematical challenge. However, one can predict some features of the process. Namely, we have analysed the mean first passage time and the mean exposure. The former can be solved exactly starting from the backward Fokker-Planck equation whereas we have estimated the latter in an approximated framework.

Regarding the mean first passage time, we have found out a good agreement between numerics and theoretical predictions for relatively low initial exposure and networks not strongly connected. Nevertheless, numerical results for a power-law out-degree distribution point out that diffusive fails to obtain a quantitatively good

prediction for the mean first passage time. In other words, when *burst* processes where the exposure changes abruptly due to the spread of the meme in users with high out-degree governs the dynamics, we cannot trust on the diffusive approximation anymore. It would be interesting to quantify and study the limits of the validity of the approximation, but to do so quantitatively is challenge beyond the purpose of this thesis.

With respect to the mean value for the exposure, we have obtained the theoretical prediction approximating the drift coefficient neglecting the quadratic term in the exposure. The final result, which is a an exponential decay, turns out to be the same one that one could obtain considering only the deterministic part of the Langevin equation, that is, neglecting the noise term. Comparison with numerical results shows up that, in spited of the rustic approximation, our estimation works quite well.

In chapter 3, motivated by the aim of applying the control principles on a real-world network, we have studied a sample of videos from YouTube collected in the period of General Italian Election of 2018 about political leaders. By mean of a web crawler, as described in section 3.1, we have collected and analysed in time a number of videos related to the principal political leaders. In particular, a natural tool for studying the collected data is Network Theory, which have allowed us to focus not only on the dynamics of the number of views per video, but also on the recommendation system between them. Indeed, we have carried out a centrality analysis for investigating possible correlations between this intrinsic structural properties of the graphs and the number of views, our final goal being to predict which videos were more likely to become viral. Nevertheless, although many alternative definitions of centrality were studied, including "Dynamic SpringRank" [23], we have not found out any clear pattern of correlations.

Finally, we have performed assortativity measures over the network formed by our collected videos. Interestingly, this study has shown that the YouTube recommendation system acts in some way like an "Echo chamber", inducing users to watch ideologically-aligned contents.

Future perspectives on this project are twofold: from the theoretical point of view, we would like to generalise and modify the model for memes popularity also to the YouTube network, taking in account the specific structure of the Youtube recommendation system and investigating its role in the increase or disruption of online echo chambers. On the other hand, in order to predict virality of the memes, we should introduce also semantic analysis that are typically used in computational social science, like sentimental analysis, to be able to understand the role of the memes contents in relation to their spreading.

APPENDIX A

POSSIBLE GENERALISATION FOR GLEESON MODEL

Herein, we present briefly some generalisations proposed by Gleeson et al. [28] for the model inspired by Twitter discussed in Chapter 4. This more general model attempts to mimic more realistic Twitter dynamics.

In particular, it is straightforward to generalise the model allowing each node's screen to have capacity $c \geq 1$, meaning that the screen can simultaneously hold c memes. Thus, we consider each screen to be composed of c distinct screen-slots. When retweeting, a node (user) chooses one of their c screen-slots at random to determine the meme that is transmitted to its followers. When a node innovates, or when it receives a meme from another node, the new meme is placed in a randomly-chosen screen-slot, overwriting any existing meme in that slot.

An additional generalisation is to allow for tweeted memes to be accepted onto followers' screens with probability $\lambda \leq 1$ (with $\lambda = 1$ giving the case discussed in Chapter 4).

Incorporating these generalisations into the derivation for the the following equation for PGF of the excess popularity distribution $G(a, y)$:

$$c\frac{\partial G}{\partial a} = \lambda z + \mu - (\lambda z + 1)\, G + (1 - \mu) y G f(1 - \lambda + \lambda G). \tag{A.1}$$

which reduces to Eq. (2.8) of the main text in the case $c = 1$ and $\lambda = 1$. The corresponding equation for $H(a, x)$ is $H = xGf(1 - \lambda + \lambda G)$, which can be used for finding old-age asymptotic behaviours.

A feature of the model described in the main text is that memes remain on users' screens even after the user retweets the meme, and so the same user may retweet a single meme multiple times. This is not a problem if we identify memes by hashtags.

Then, users may use the same hashtag through different messages during their activity. However, if the memes under consideration are identified as entire messages, then this feature of the model is unrealistic. Each Twitter message is usually retweeted at most once by any single user. Therefore, a possible generalisation can be simply to avoid retweeting the same meme by the same user. Nevertheless, numerical results by Gleeson et al. [28] indicate that, although the new rule becomes hard to study from the analytic point of view, the criticality mechanism is not strongly affected by the modification of the model.

APPENDIX B

THE KRAMERS-MOYAL EXPANSION

In this appendix, we briefly review the formalism of the diffusive approximation for a general Master Equation.

We consider a Master Equation for a single[1] discrete random variable x as the one in Eq. (2.24), i.e.

$$\partial_t P(x,t) = \sum_{x'} w(x|x',t)P(x',t) - w(x'|x,t)P(x,t), \tag{B.1}$$

where $w(x|x',t)$ is the transition rate of the process leading from $x' \to x$ at time t. Note that, we have written the master equation for the probability density function $P(x,t)$ instead that for the conditional PDF $P(x,t|x_0,0)$ as in the main text. We recall that both probabilities fulfill the same master equation, being the initial condition the only difference.

Let us suppose that transition rates $w(x'|x,t) \neq 0$ only for $x - x' \equiv \Delta x \in E$, where E is a finite set. Now, we define $\hat{W}(\Delta x|x',t) \equiv w(x|x',t)$. The so called Kramer-Moyal expansion is simply a Taylor expansion of (B.1) around $x - \Delta x = x$, that is,

$$\begin{aligned} \partial_t P(x,t) &= \sum_{\Delta x} \hat{W}(\Delta x|x-\Delta x,t)P(x-\Delta x,t) - \hat{W}(\Delta x|x,t)P(x,t) \\ &= \sum_{k=1}^{+\infty} \frac{(-1)^k}{k!} \partial_x^k \left[A^{(k)}(x,t)P(x,t) \right], \end{aligned} \tag{B.2}$$

[1] The generalisation to the case of many variable follows straightforwardly.

where

$$A^{(k)}(x,t) = \sum_{\Delta x \in E} \hat{W}(\Delta x|x,t)\Delta x^k, \tag{B.3}$$

which is usually called as the k-th jump moment of the process.

As explained by Gardiner [26], if we keep only the first term in Eq. (B.2) we get a deterministic motion whose trajectories are regulated by

$$\dot{x}(t) = A^{(1)}(x(t),t). \tag{B.4}$$

Truncating at second order, we end up with the diffusive approximation, which is the Fokker-Planck equation we have work with though the main text,

$$\partial_t P(x,t) = -\left[A^{(1)}(x,t)P(x,t)\right] + \frac{1}{2}\left[A^{2)}(x,t)P(x,t)\right]. \tag{B.5}$$

Note that the drift and diffusion coefficients $A(x)$ and $B(x)$ in the main text are exactly the first a second jump moment respectively.

Remarkably, truncating Eq. (B.2) at any order greater that 2 in general might lead to negative probabilities whereas the first two terms guarantee that this does not happen. Precisely, there is a theorem by Pawula which states that the expansion either stops after the first term or the second term. Otherwise, if the expansion continues beyond the second term, it must contain an infinite number of terms in order that the solution to the equation be interpretable as a probability density function.

APPENDIX C

BACKWARD FOKKER-PLANCK EQUATION

In this appendix, we put forward the basics of the backward Fokker-Planck equation. Our starting point will be the usual Forward Fokker-Planck Equation

$$\partial_t P(x,t|x_0,t_0) = -\partial_x A(x) P(x,t|x_0,t_0) + \frac{1}{2}\partial_x^2 B(x) P(x,t|x_0,t_0). \tag{C.1}$$

This give us the evolution $P(x,t|x_0,t_0)$ with t. Conversely, one may be interested in the evolution with t_0, this is the purpose of the backward Fokker-Planck equation.

Let us consider a system whose dynamics is regulated by Eq. (C.1). Since it is Markovian, the joint probability to find the system at x_0, x' and x at successive times $t_0 < t' < t$ reads

$$P(x,t;x',t';x_0,t_0) = P(x,t|x',t')P(x',t'|x_0,t_0)P(x_0,t_0), \tag{C.2}$$

where of course $P(x0,t0)$ is the probability that the system was at x_0 at t_0. Simply integrating over x' this equation, from the left hand side we get

$$\int dx' P(x,t;x',t';x_0,t_0) = P(x,t;x_0,t_0) = P(x,t|x_0,t_0)P(x_0,t_0), \tag{C.3}$$

while the right hand side reads

$$\int dx' P(x,t|x',t')P(x',t'|x_0,t_0)P(x_0,t_0). \tag{C.4}$$

Therefore, equating both terms we get that $P(x,t|x_0,t_0)$ is the solution of the Chapman-Kolmogorv equation

$$P(x,t|x_0,t_0) = \int dx' P(x,t|x',t')P(x',t'|x_0,t_0) \quad \forall\, t' \in (t_0,t). \tag{C.5}$$

Taking the derivative of this equation with respect to t' the left hand side is 0, because it does not depend on t', and therefore

$$0 = \int dx' \Big\{ \underbrace{\left[\partial_{t'} P(x,t|x',t')\right] P(x',t'|x_0,t_0)}_{I^{(1)}} + \underbrace{P(x,t|x',t')\partial_{t'} P(x',t'|x_0,t_0)}_{I^{(2)}} \Big\}. \tag{C.6}$$

If we now take the limit $t' \to t_0$, remembering that $\lim_{t'\to t_0} P(x',t'|x_0,t_0) = \delta(x'-x_0)$ we get

$$\int dx' I^{(1)} = \partial_{t_0} P(x,t|x_0,t_0) \tag{C.7a}$$

$$\int dx' I^{(2)} = A(x_0)\partial_{x_0} P(x,t|x_0,t_0) + \frac{1}{2} B(x_0)\partial^2_{x_0} P(x,t|x_0,t_0) \tag{C.7b}$$

where we have integrated by parts to obtain the integral of $I^{(2)}$. Therefore,

$$\partial_{t_0} P(x,t|x_0,t_0) = -A(x_0)\partial_{x_0} P(x,t|x_0,t_0) - \frac{1}{2} B(x_0)\partial^2_{x_0} P(x,t|x_0,t_0) \tag{C.8}$$

Since the dynamics is homogenous in time, i.e. $P(x,t|x_0,t_0) = P(x,t+\tau|x_0,t_0+\tau)$, we have the following relation

$$\partial_t P(x,t|x_0,t_0) = -\partial_{t_0} P(x,t|x_0,t_0). \tag{C.9}$$

Finally, applying this symmetry to equation (C.8) leads to the Backward Fokker-Plank equation

$$\partial_t P(x,t|x_0,t_0) = A(x_0)\partial_{x_0} P(x,t|x_0,t_0) + \frac{1}{2} B(x_0)\partial^2_{x_0} P(x,t|x_0,t_0). \tag{C.10}$$

APPENDIX D

NUMERICAL SIMULATIONS RESULTS

In this appendix, we present some results of the numerical simulations described in section B.3 of the main text. In particular, Fig. D.1shows MFPT $T(x)$ versus number of trajectories simulated for computing it for x ranging from 0.0 to 0.55 for the process described by Langevin equation (2.46). Figures D.2 and D.3report some of the trajectories of exposure x for simulation (C) and (D) of the main text.

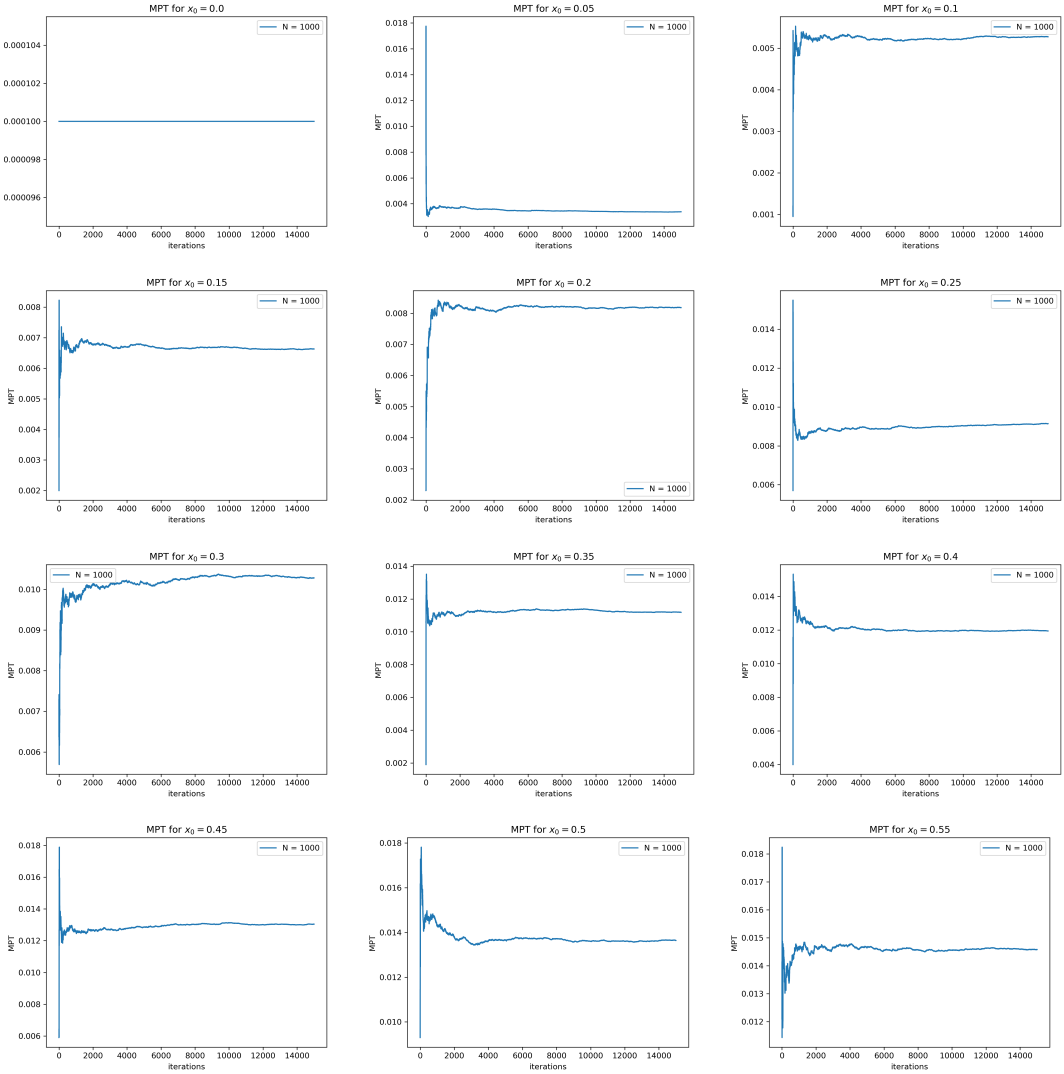


FIGURE D.1: Mean first passage time $T(x)$ versus number of trajectories simulated for computing it for x ranging from 0.0 to 0.55 for the process described by Langevin equation (2.46).

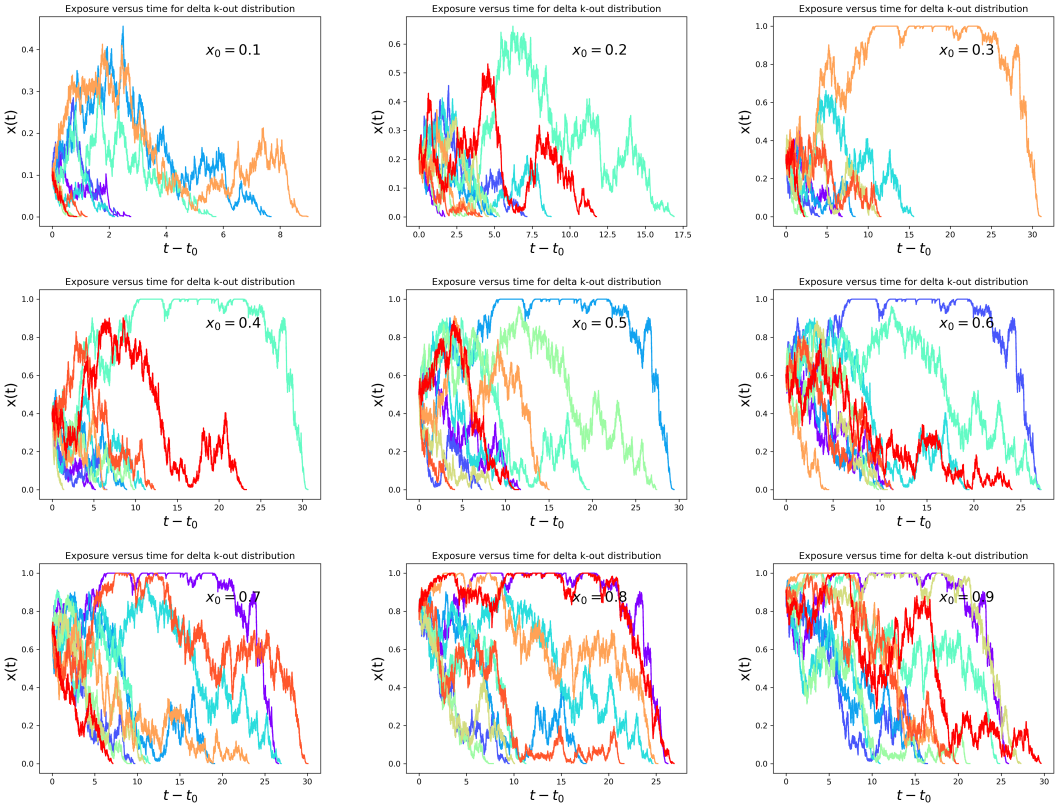


FIGURE D.2: Some trajectories for different x for simulation (C), namely for $p_k = \delta_{k,10}$ and $\tilde{\mu} = 1$.

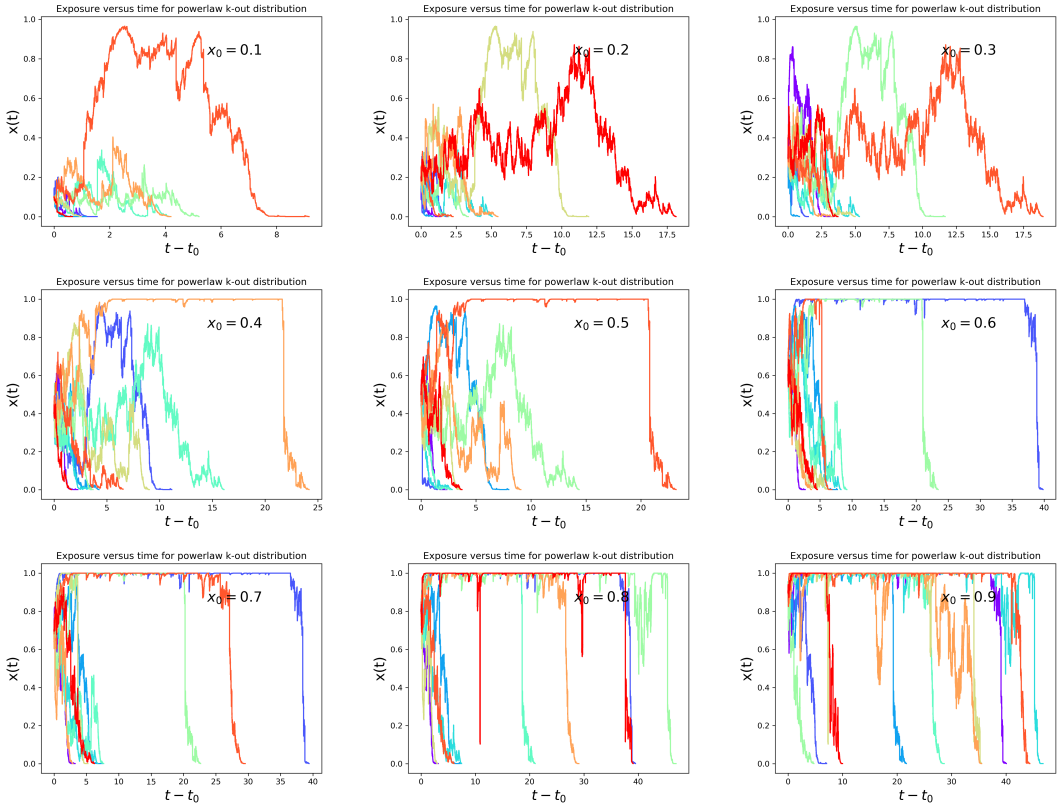


FIGURE D.3: Some trajectories for different x for simulation (D), namely for $p_k \propto k^{-2.5}$ and $\tilde{\mu} = 1$.

APPENDIX E

DATA SAMPLE FROM YOUTUBE CRAWLER

FIGURE E.1: Out-degree unified distributions of videos collected on days 0,1,2,3,4 and 5.

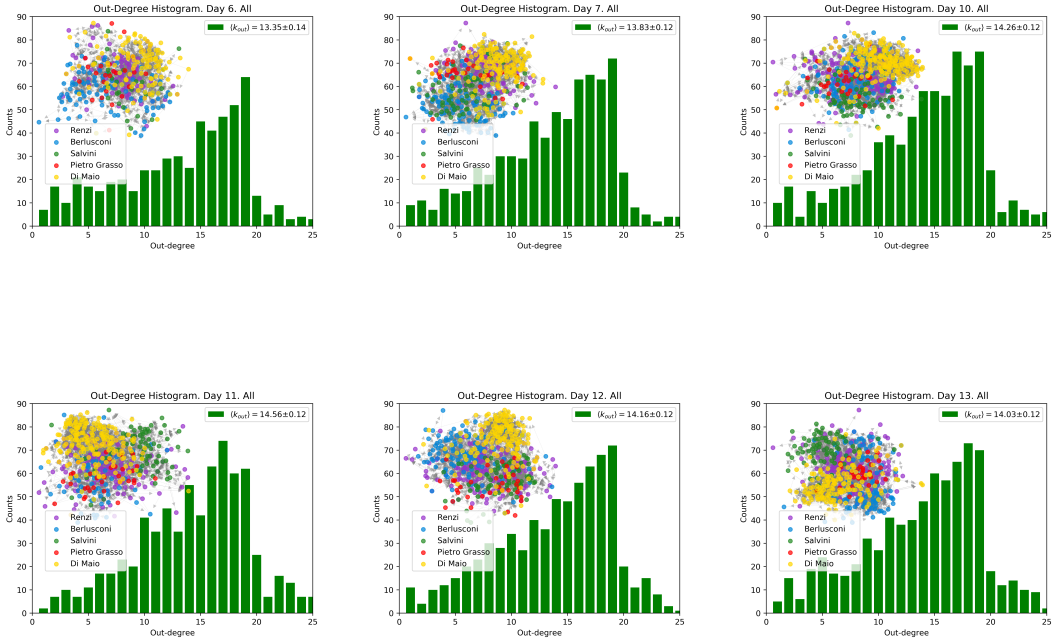


FIGURE E.2: Out-degree unified distributions of videos collected on days 6,7,10,11,12 and 13.

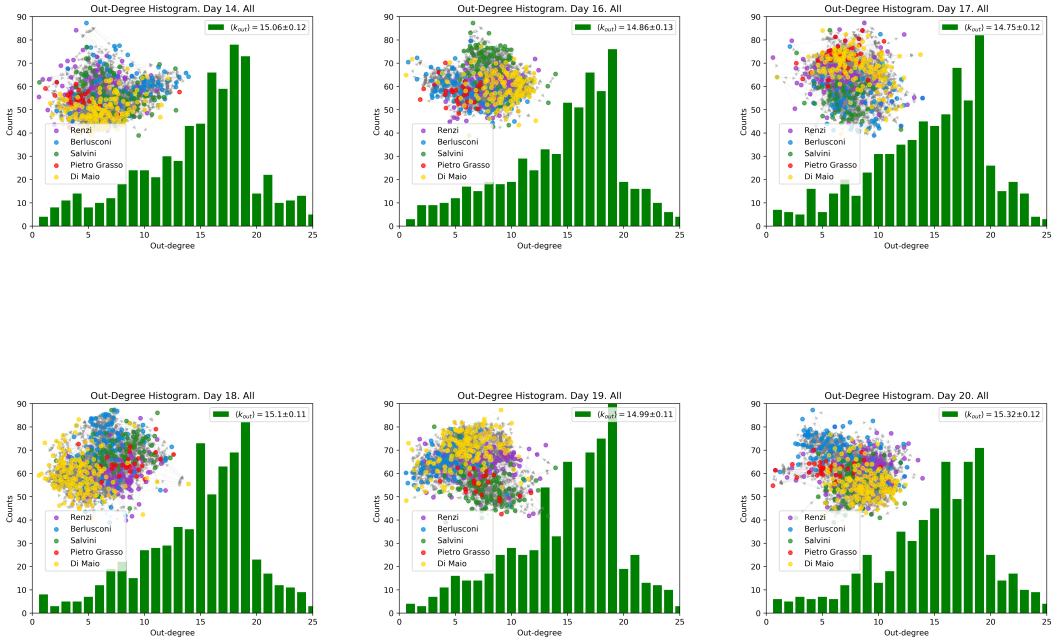


FIGURE E.3: Out-degree unified distributions of videos collected on days 14,16,17,18,19 and 20.

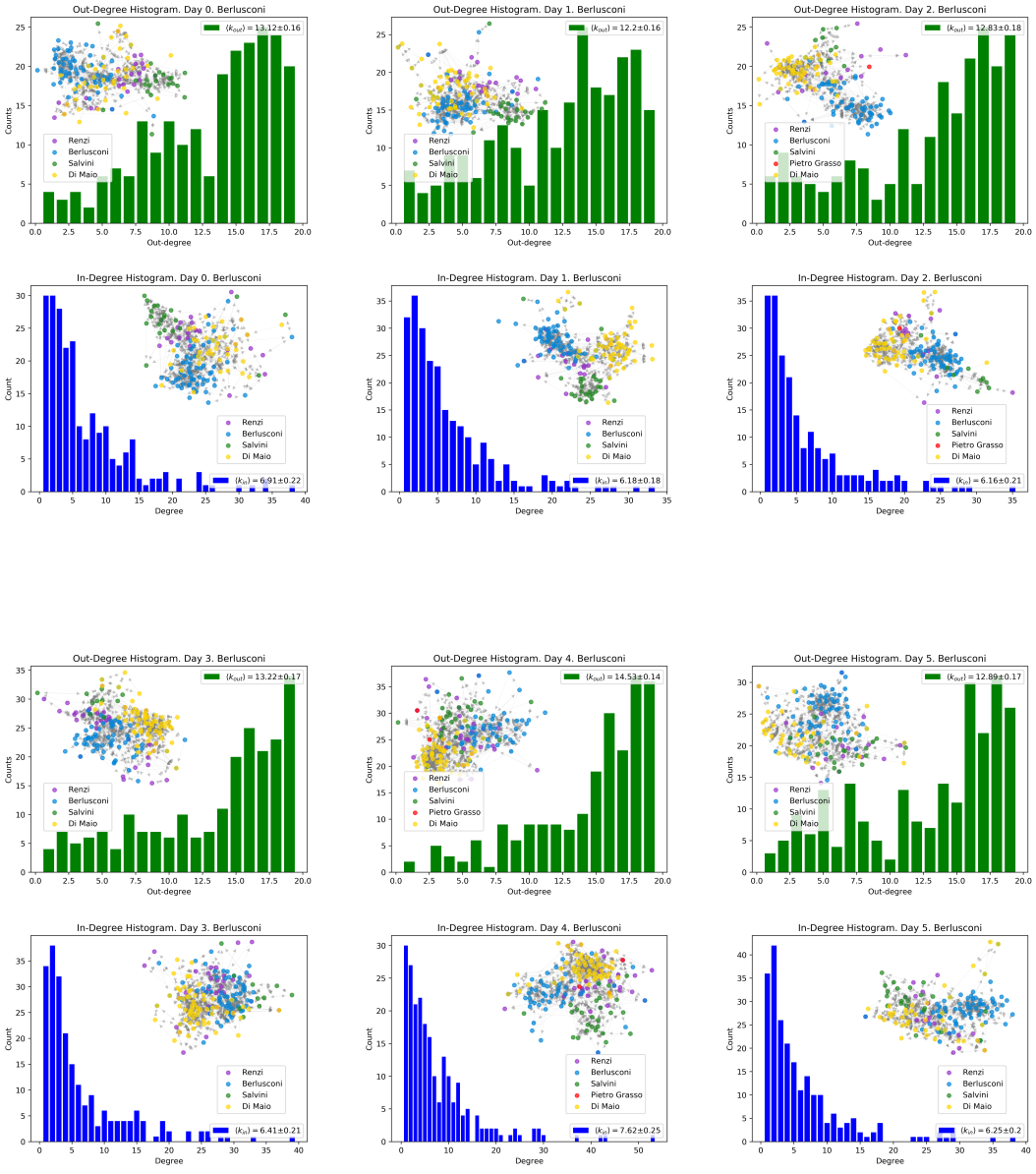


FIGURE E.4: In-degree and out-degree distributions of videos collected for "Berlusconi" as initial keyword on days 0,1,2,3,4 and 5.

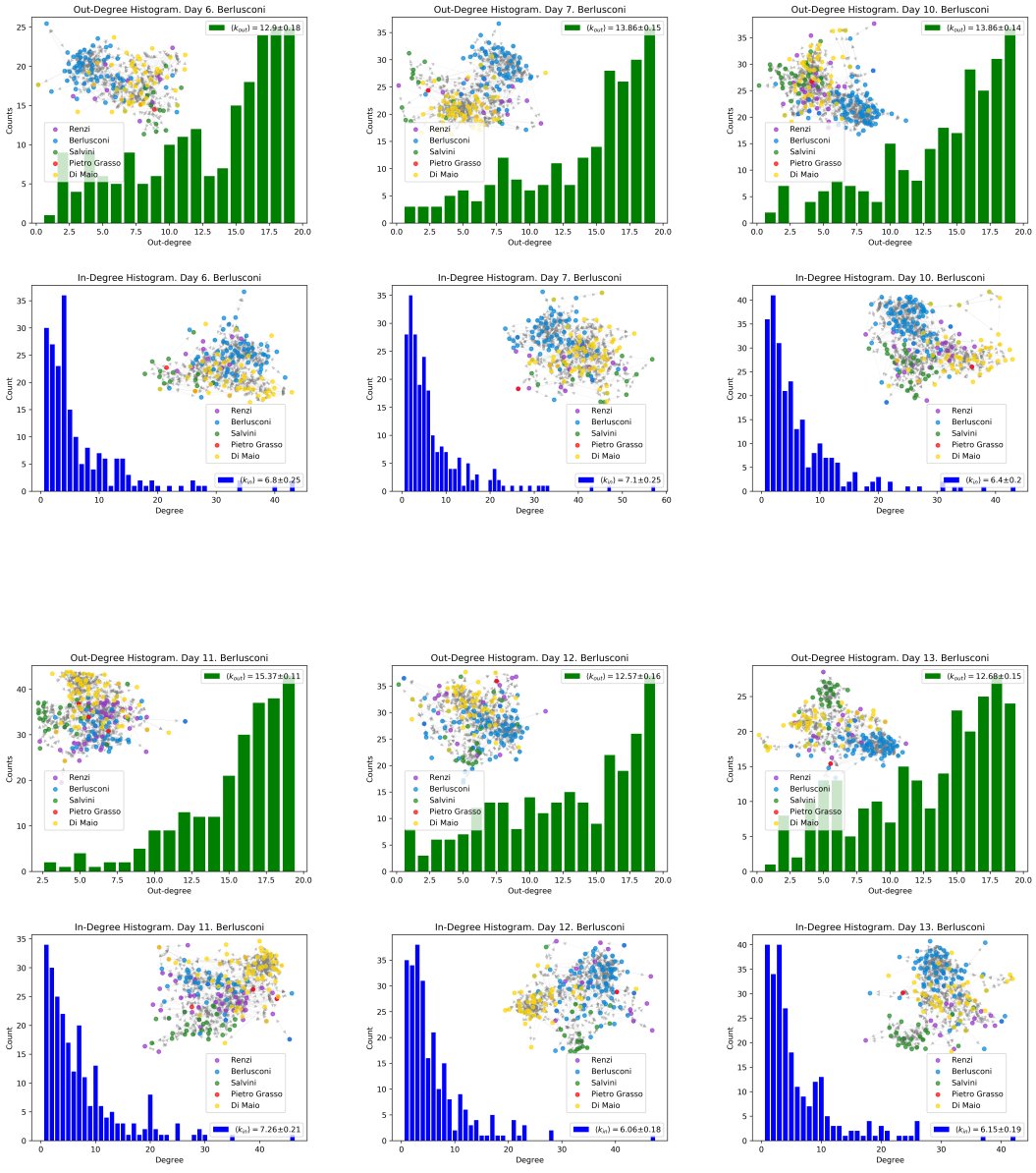


FIGURE E.5: In-degree and out-degree distributions of videos collected for "Berlusconi" as initial keyword on days 6,7,10,11,12 and 13.

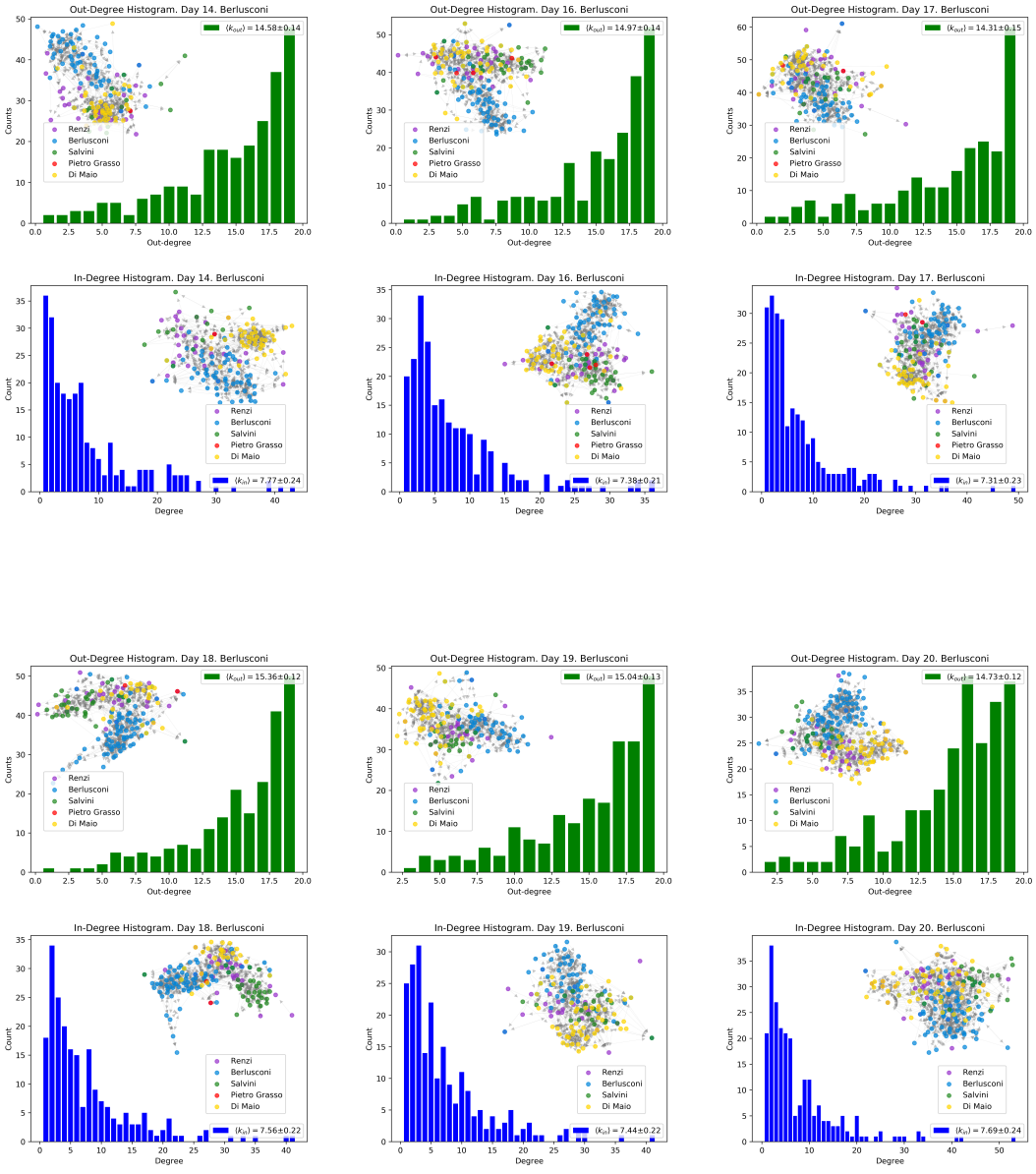


FIGURE E.6: In-degree and out-degree distributions of videos collected for "Berlusconi" as initial keyword on days 14,16,17,18,19 and 20.

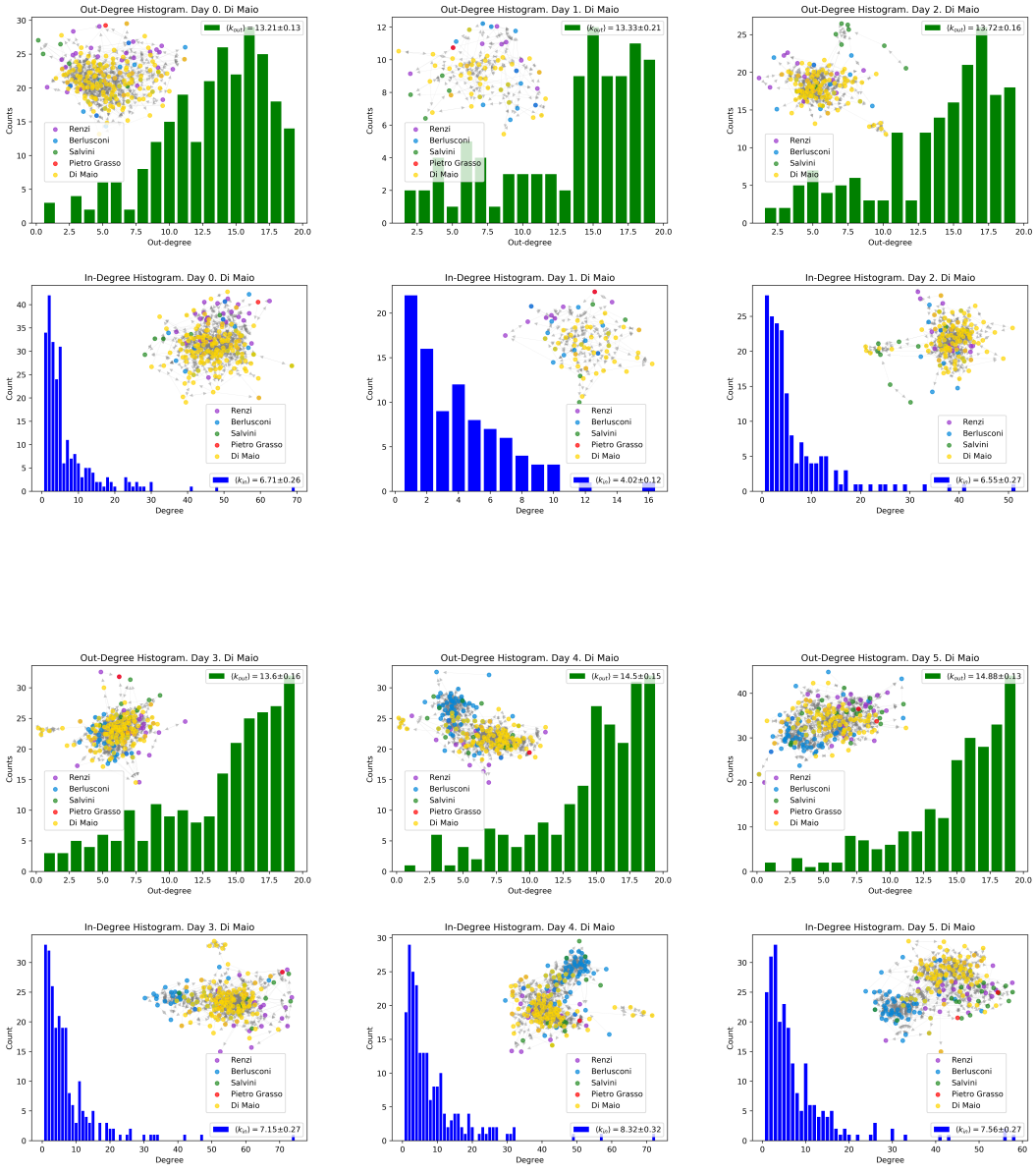


FIGURE E.7: In-degree and out-degree distributions of videos collected for "Di Maio" as initial keyword on days 0,1,2,3,4 and 5.

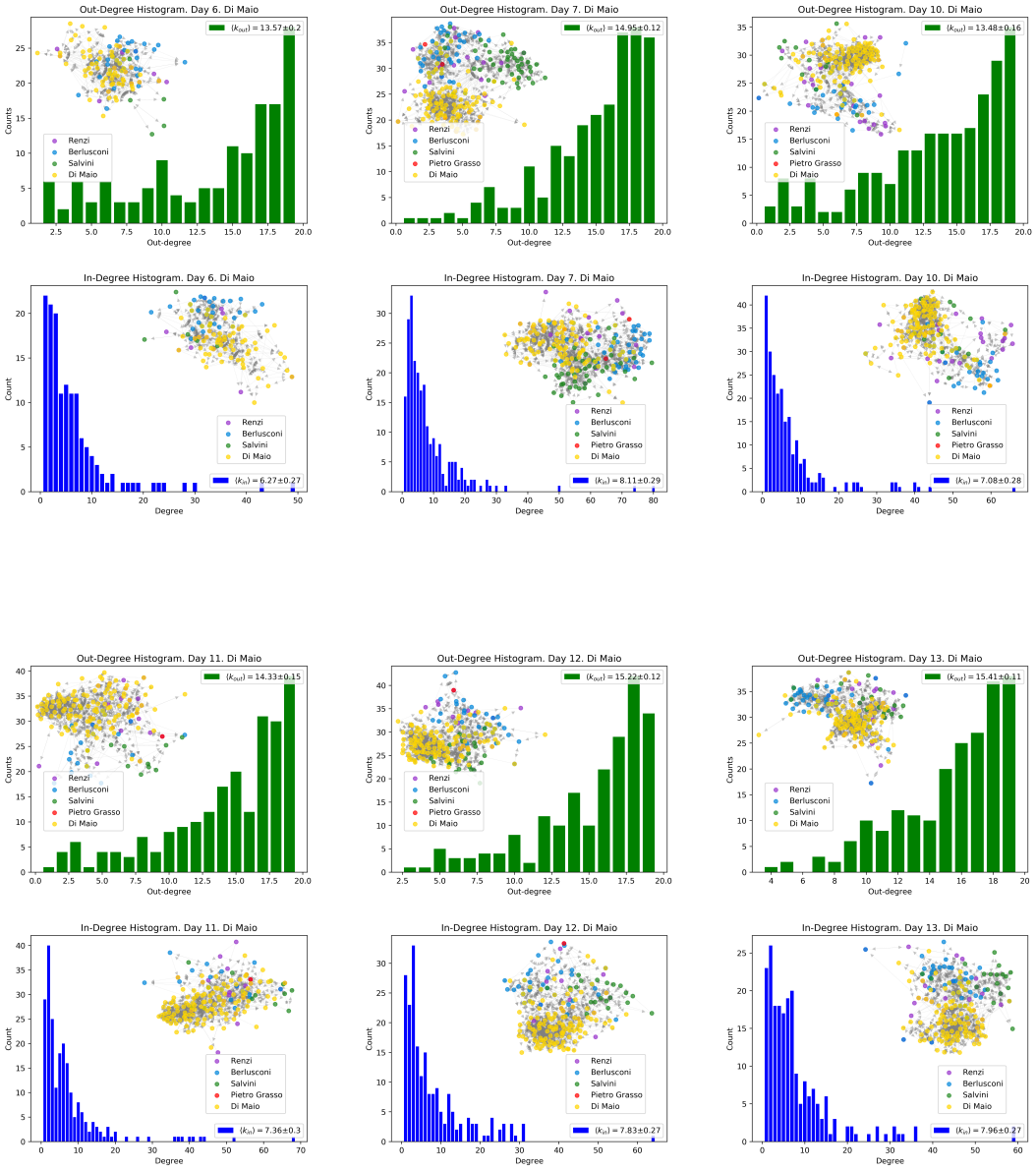


FIGURE E.8: In-degree and out-degree distributions of videos collected for "Di Maio" as initial keyword on days 6,7,10,11,12 and 13.

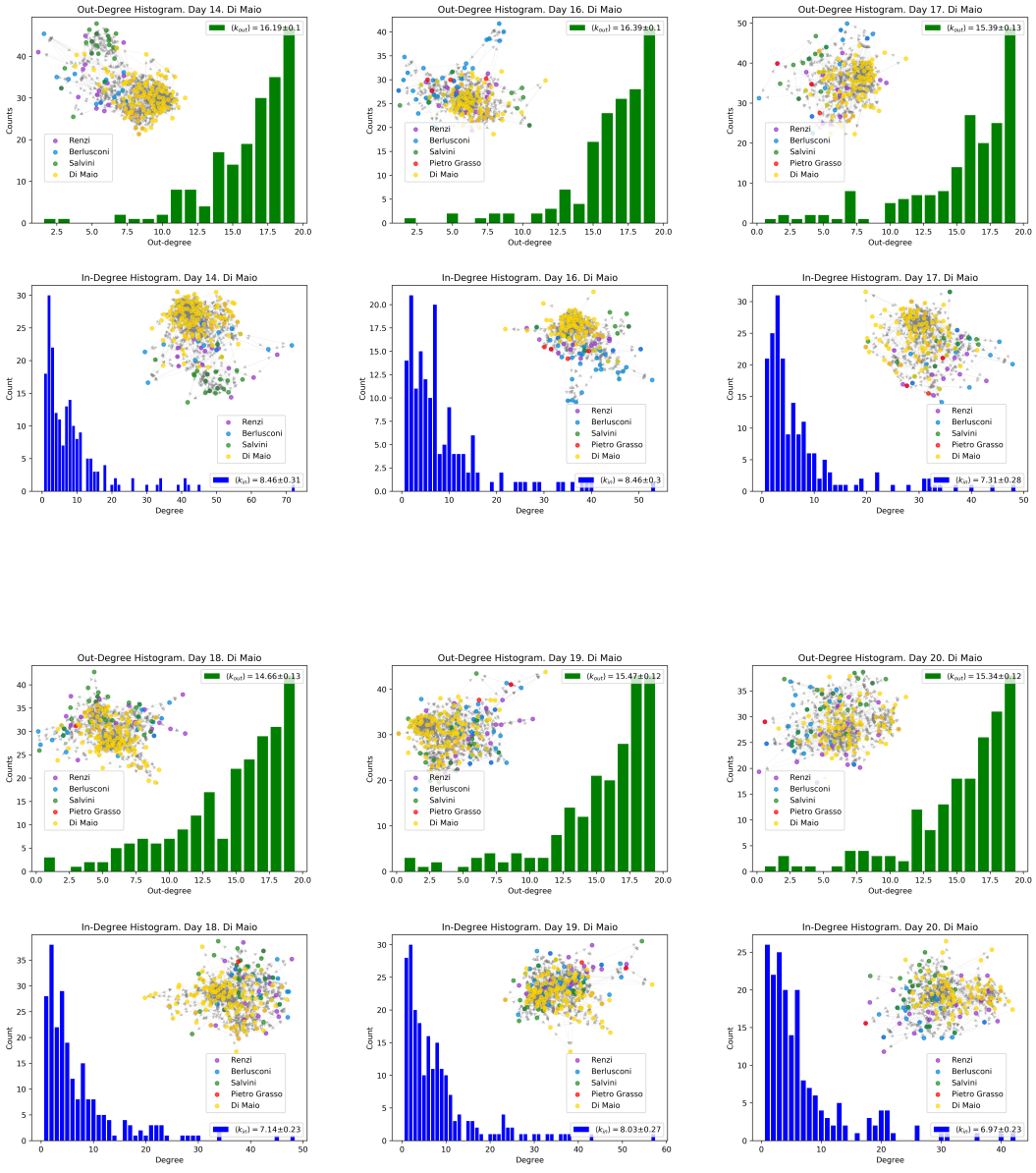


FIGURE E.9: In-degree and out-degree distributions of videos collected for "Di Maio" as initial keyword on days 14,16,17,18,19 and 20.

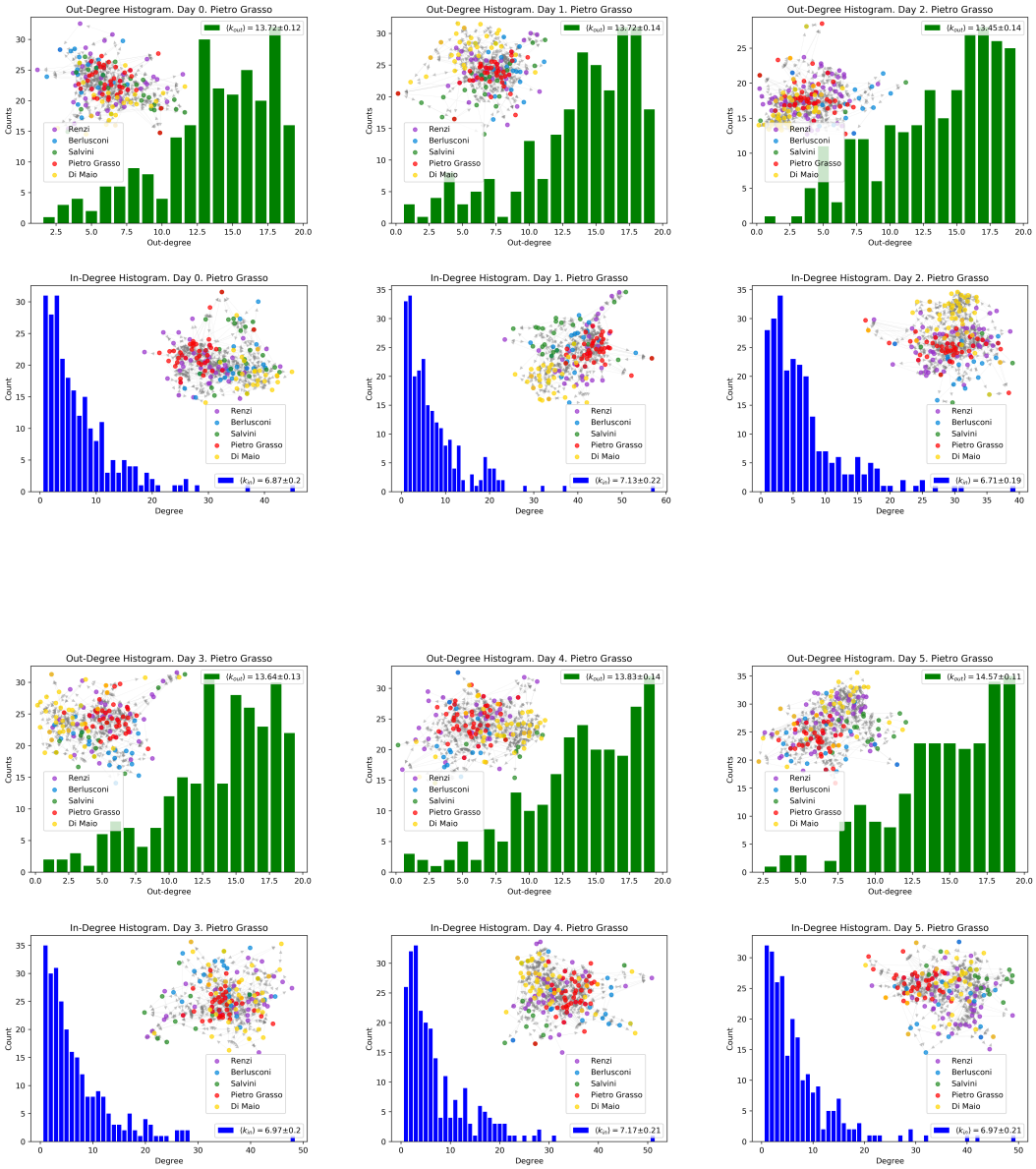


FIGURE E.10: In-degree and out-degree distributions of videos collected for "Pietro Grasso" as initial keyword on days 0,1,2,3,4 and 5.

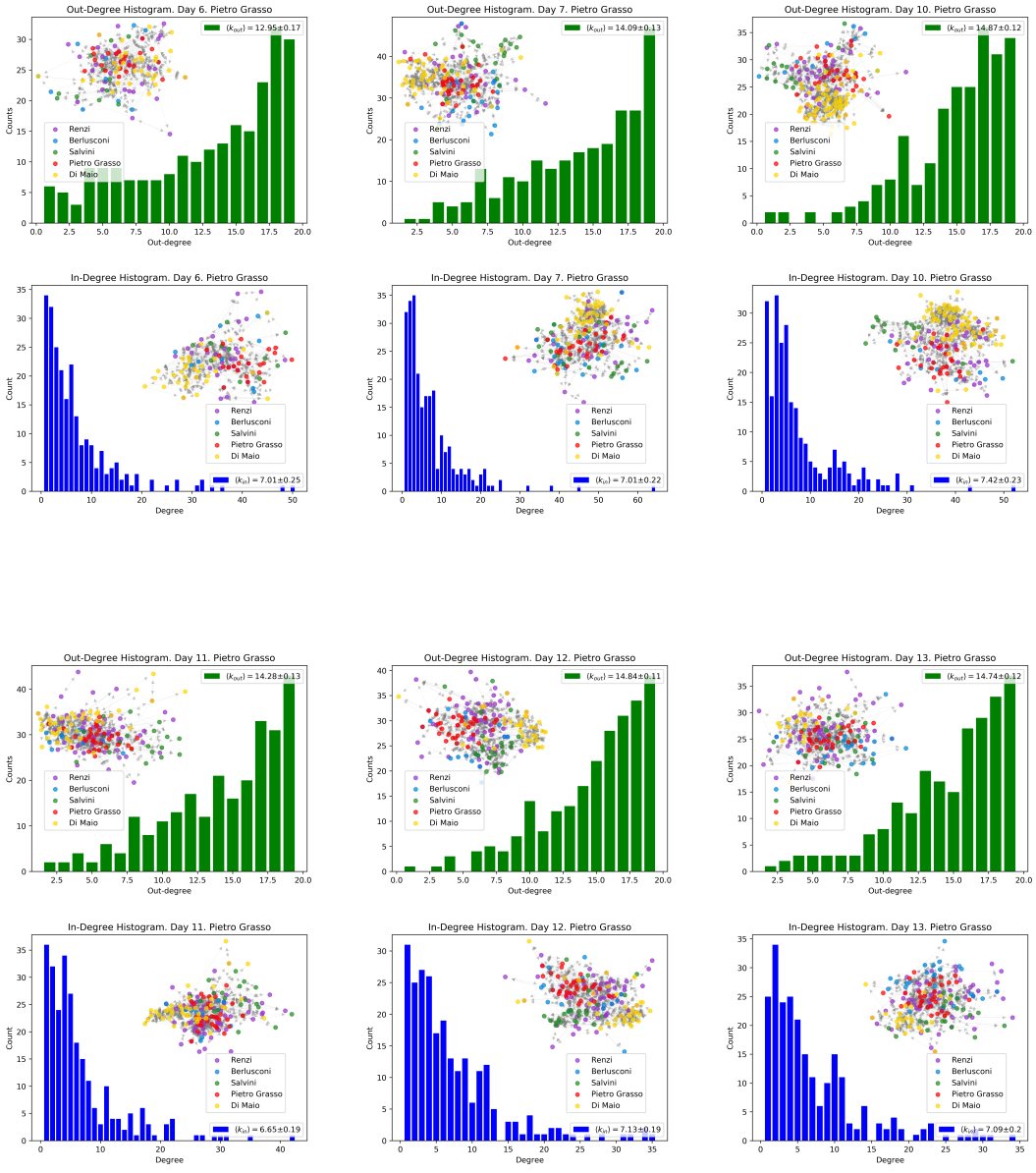


FIGURE E.11: In-degree and out-degree distributions of videos collected for "Pietro Grasso" as initial keyword on days 6,7,10,11,12 and 13.

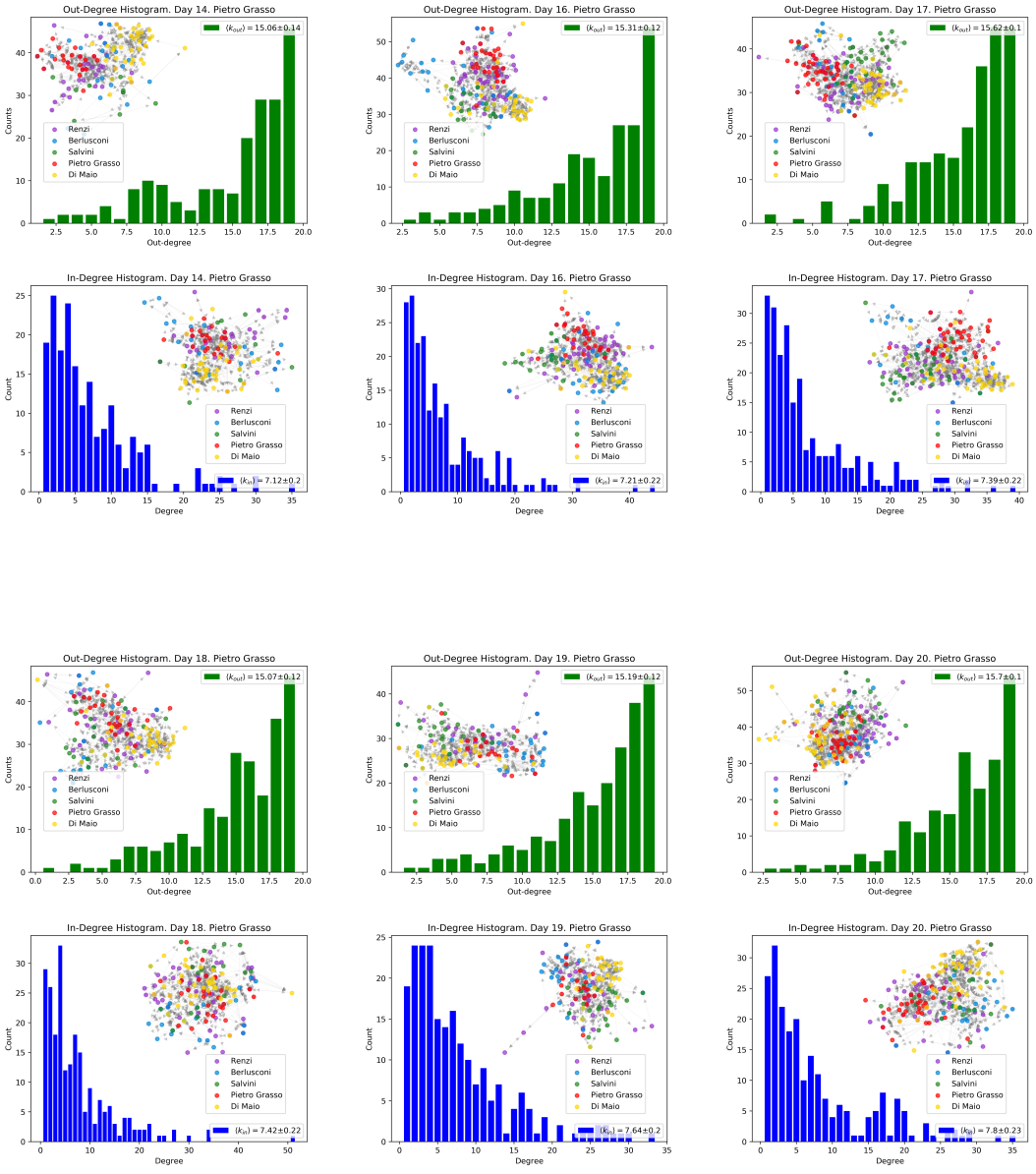


FIGURE E.12: In-degree and out-degree distributions of videos collected for "Pietro Grasso" as initial keyword on days 14,16,17,18,19 and 20.

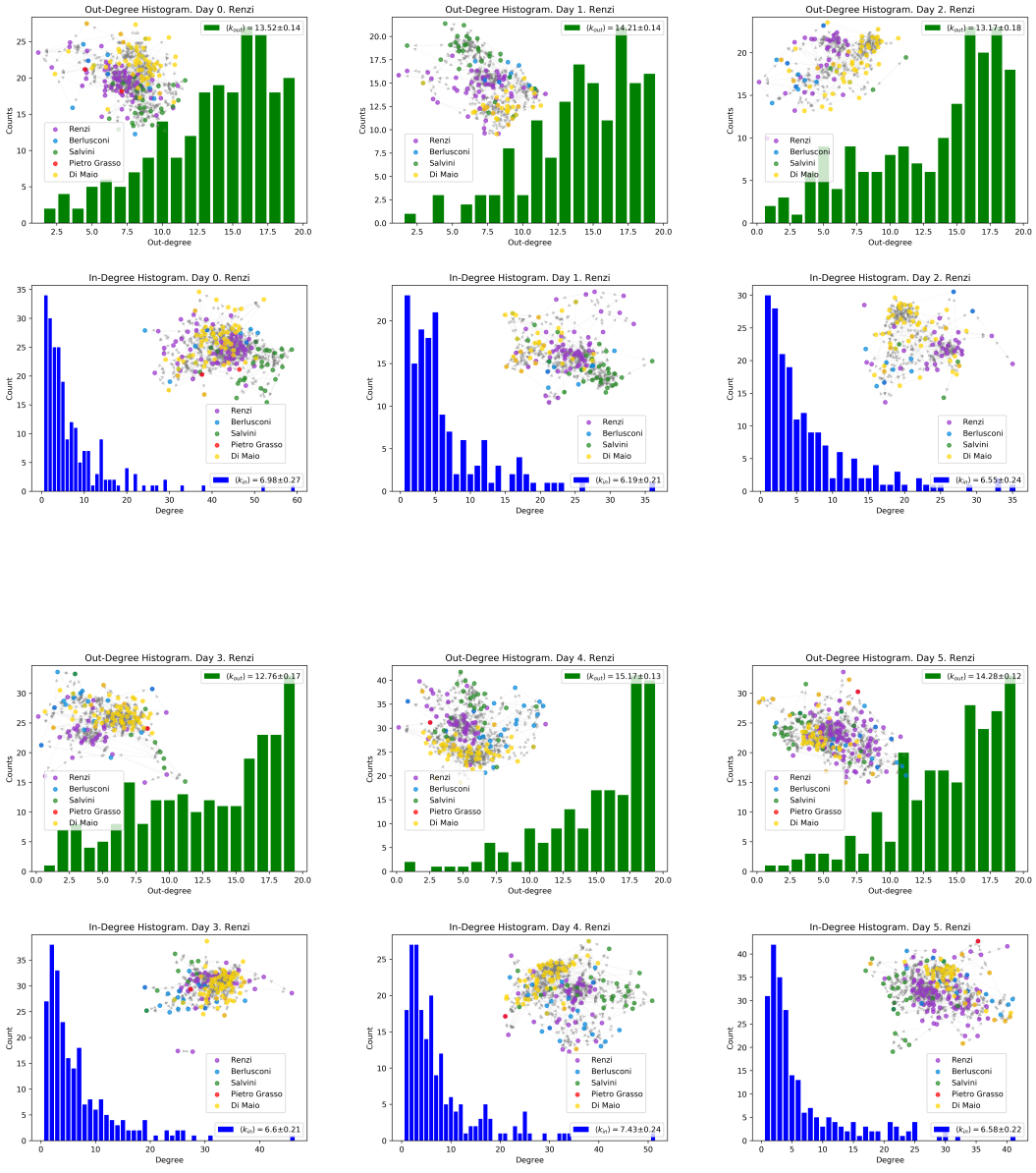


FIGURE E.13: In-degree and out-degree distributions of videos collected for "Renzi" as initial keyword on days 0,1,2,3,4 and 5.

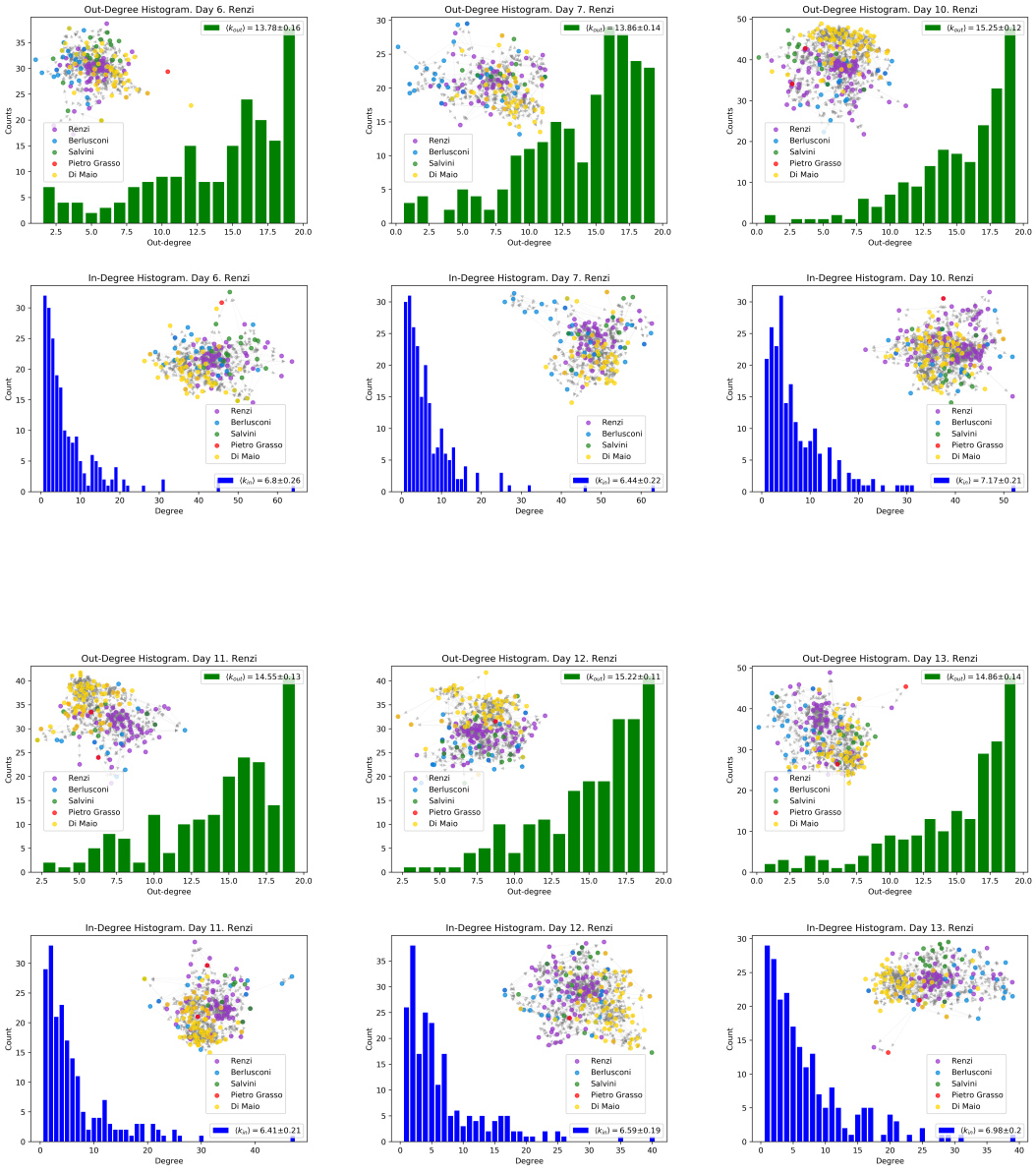


FIGURE E.14: In-degree and out-degree distributions of videos collected for "Renzi" as initial keyword on days 6,7,10,11,12 and 13.

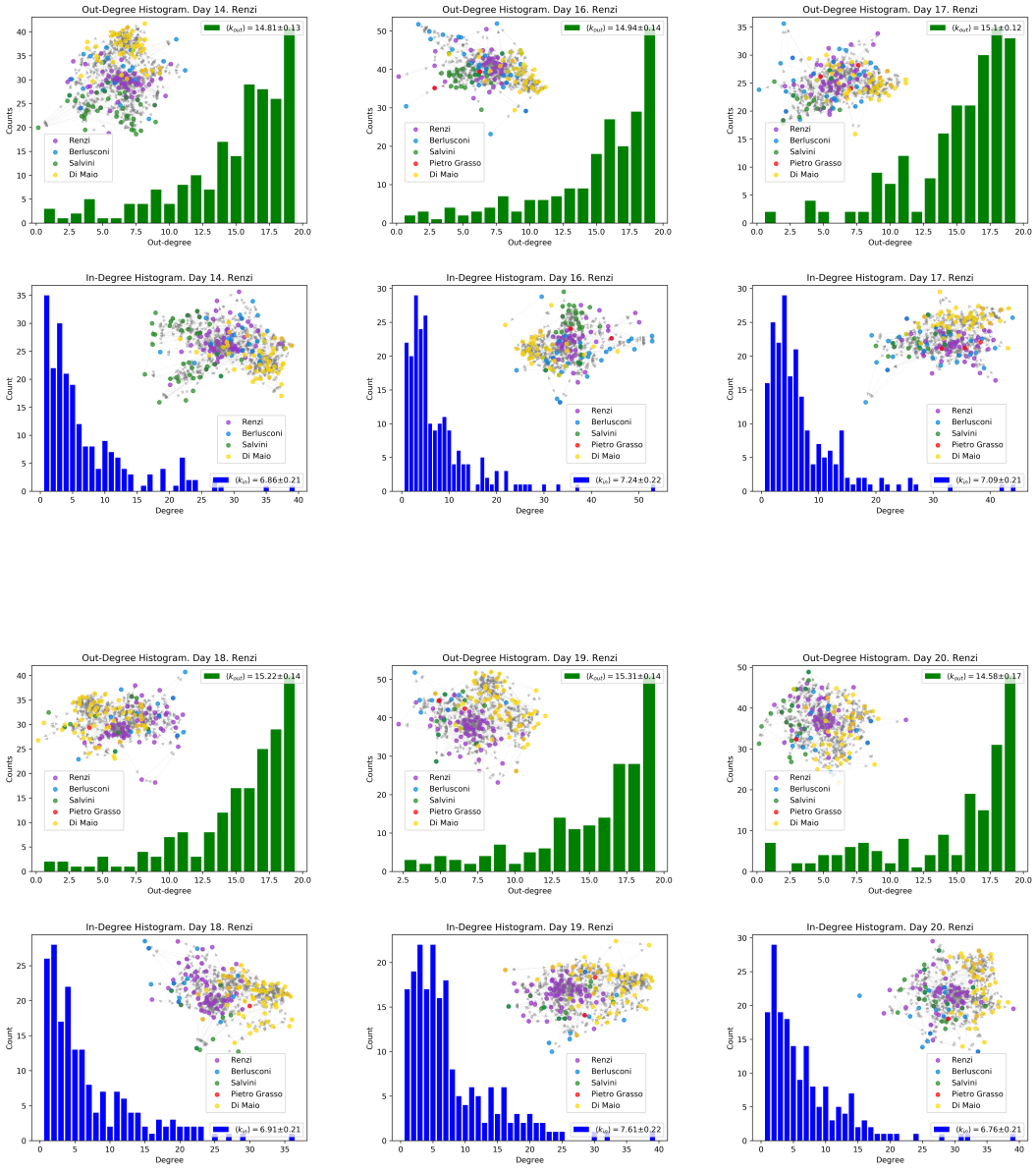


FIGURE E.15: In-degree and out-degree distributions of videos collected for "Renzi" as initial keyword on days 14,16,17,18,19 and 20.

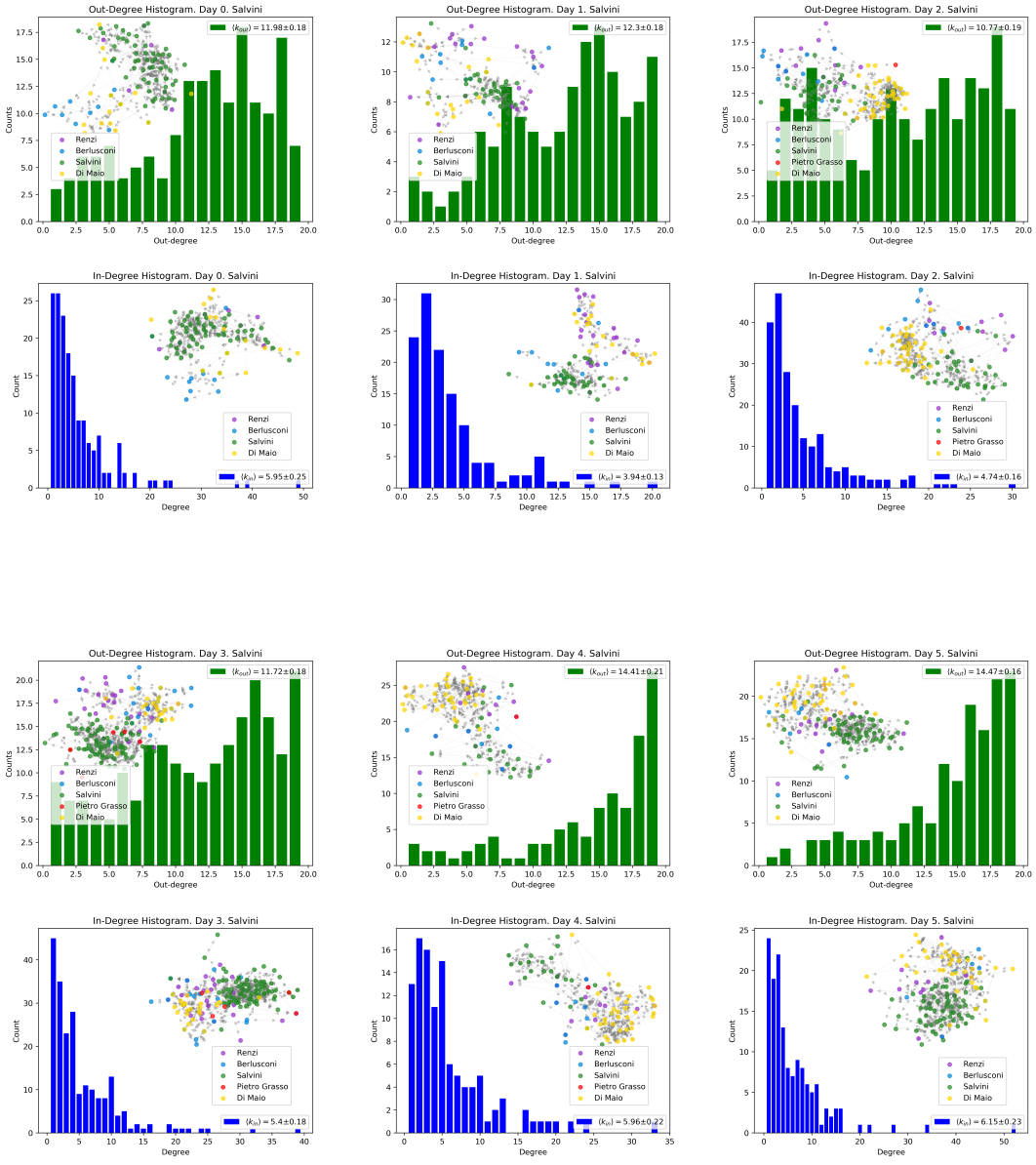


FIGURE E.16: In-degree and out-degree distributions of videos collected for "Salvini" as initial keyword on days 0,1,2,3,4 and 5.

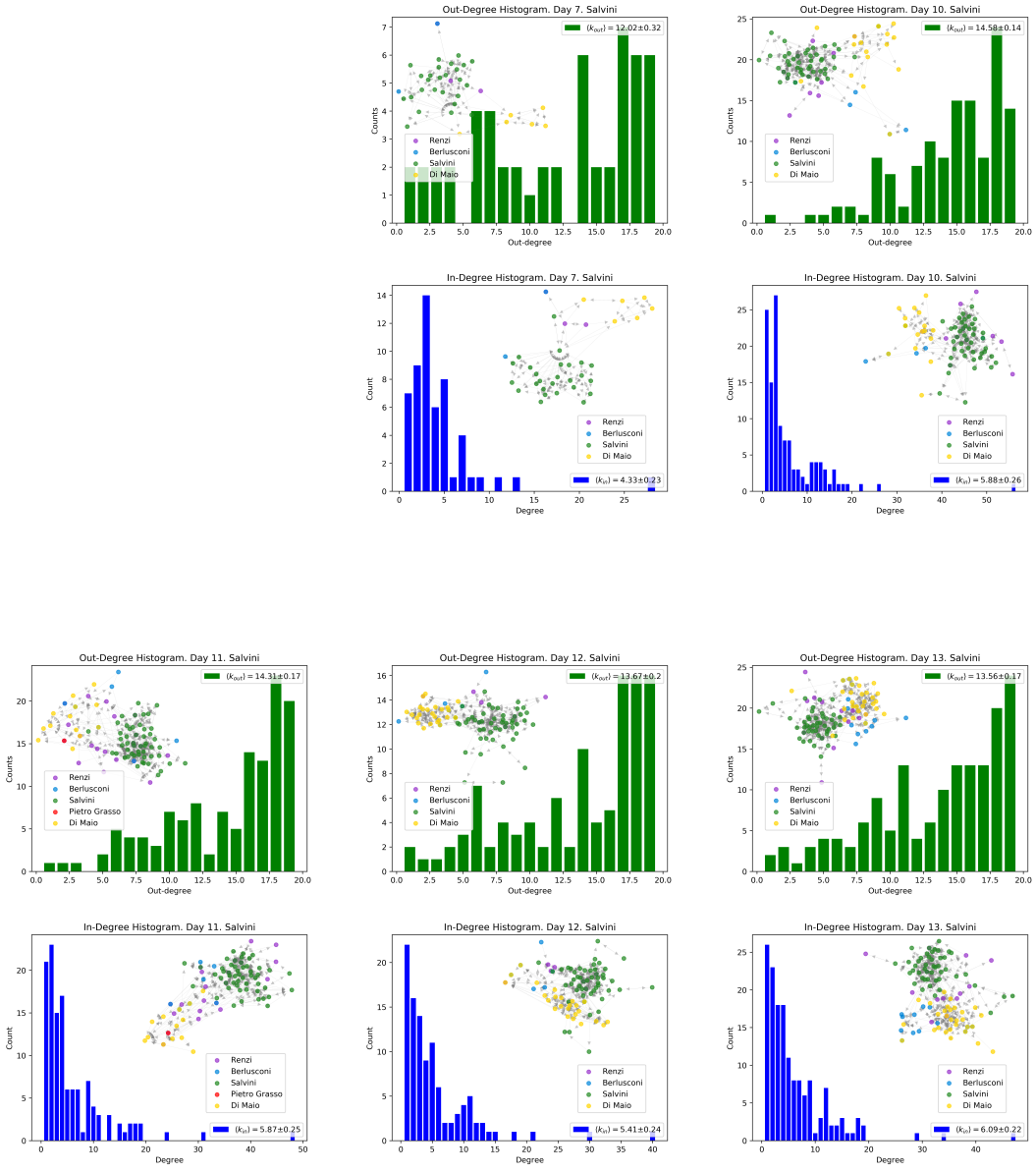


FIGURE E.17: In-degree and out-degree distributions of videos collected for "Salvini" as initial keyword on days 7,10,11,12 and 13. Day 6 is missing as previously pointed out.

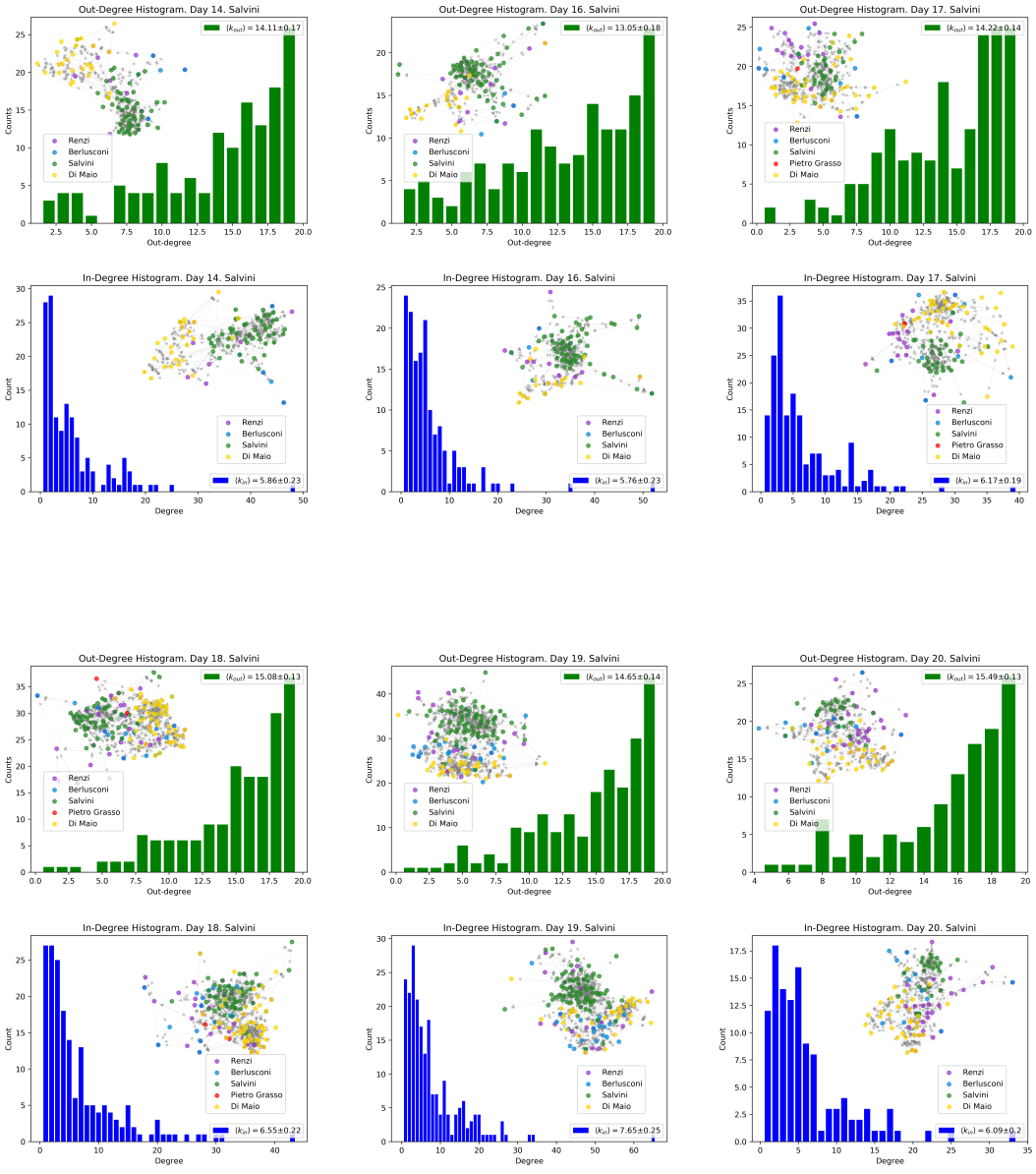


FIGURE E.18: In-degree and out-degree distributions of videos collected for "Salvini" as initial keyword on days 14,16,17,18,19 and 20.

APPENDIX F

CENTRALITY ANALYSIS OF COLLECTED VIDEOS

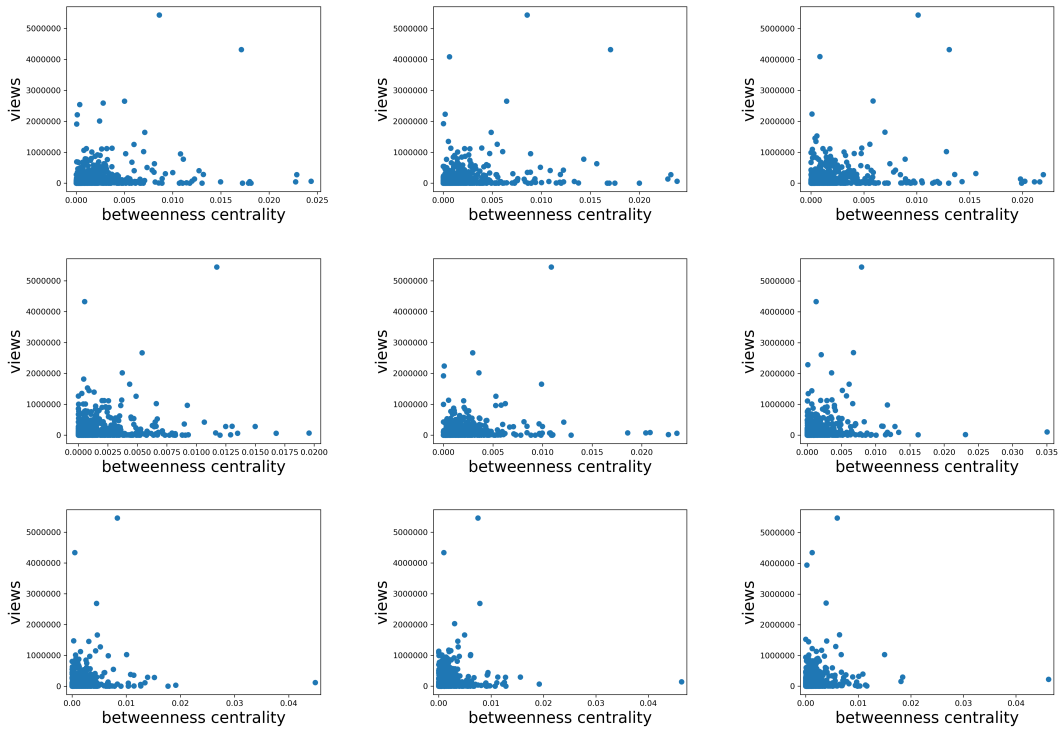


FIGURE F.1: Number of views versus betweenness centralities. Plots refer videos collected on days 0,1,2,3,4,5,6,7 and 10.

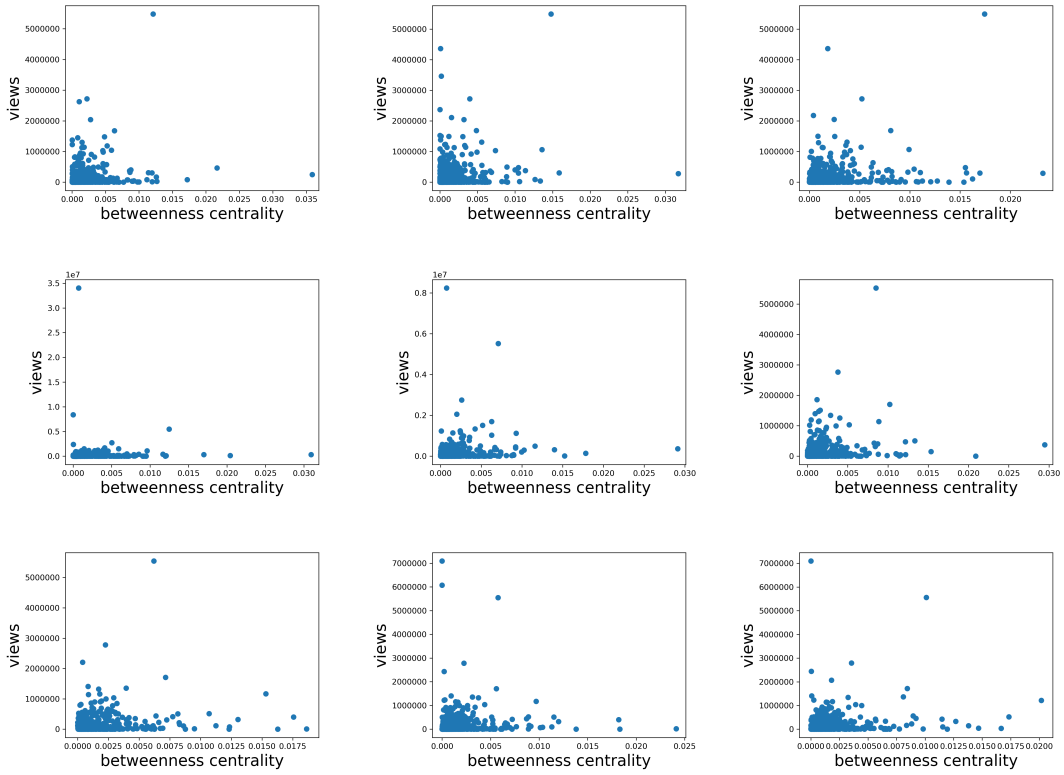


FIGURE F.2: Number of views versus betweenness centralities. Plots refer videos collected on days 11,12,13,14,16,17,18,19 and 20.

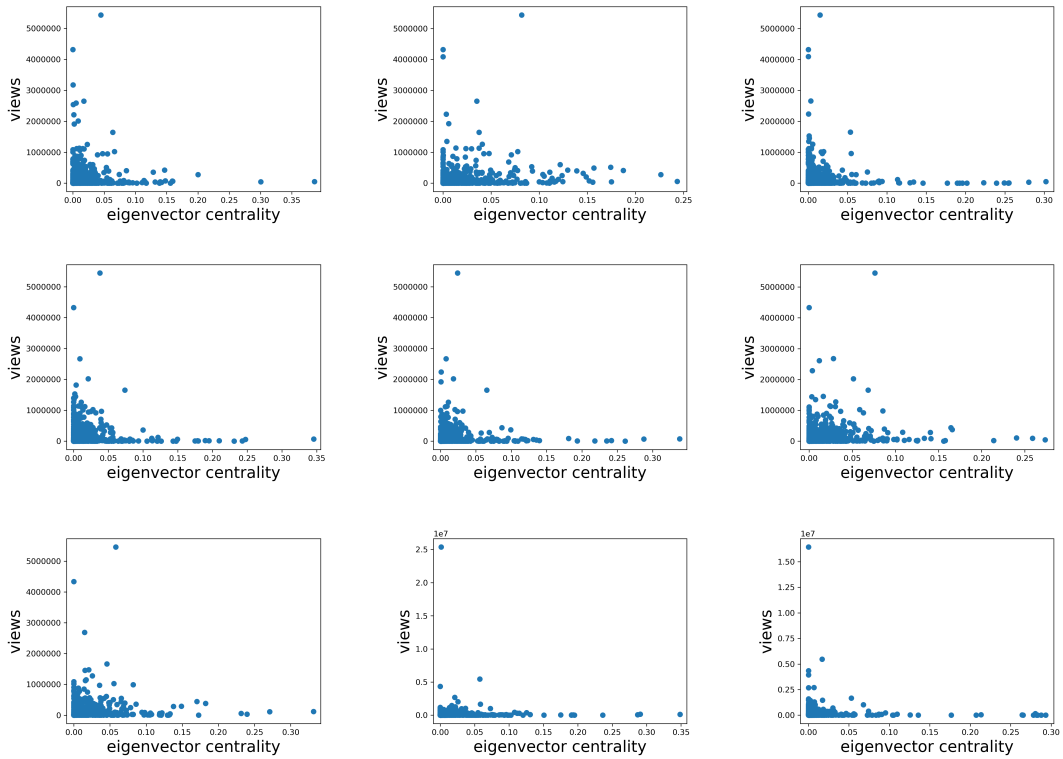


FIGURE F.3: Number of views versus eigenvector centralities. Plots refer videos collected on days 0,1,2,3,4,5,6,7 and 10.

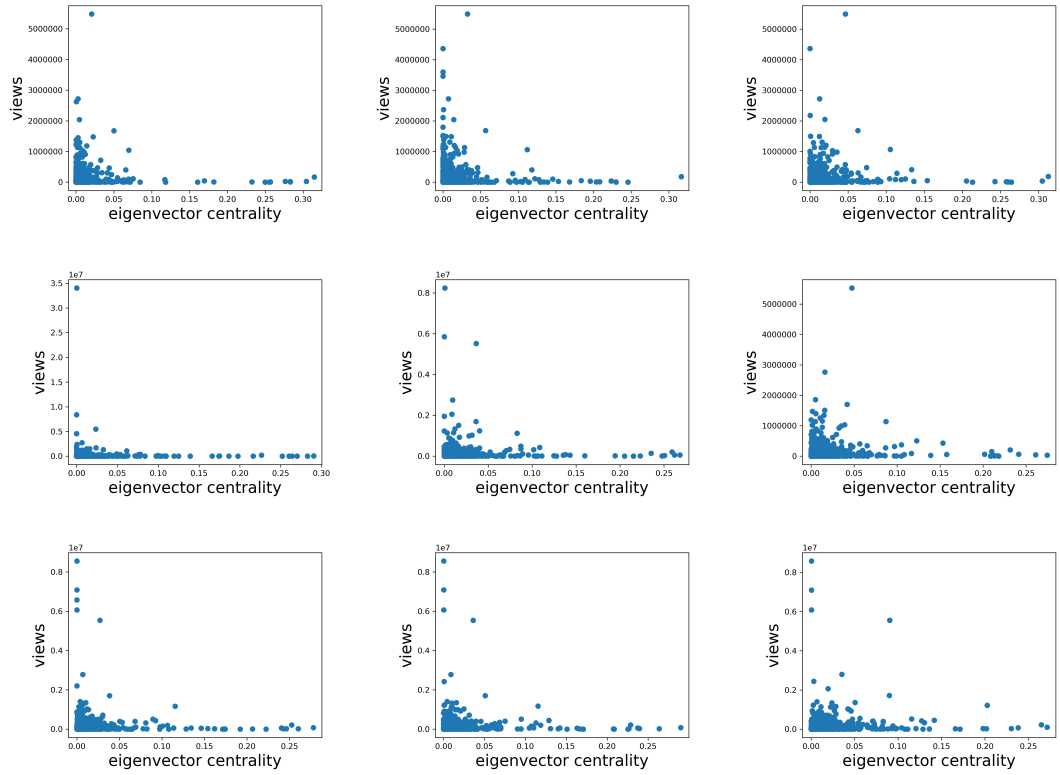


FIGURE F.4: Number of views versus eigenvector centralities. Plots refer videos collected on days 11,12,13,14,16,17,18,19 and 20.

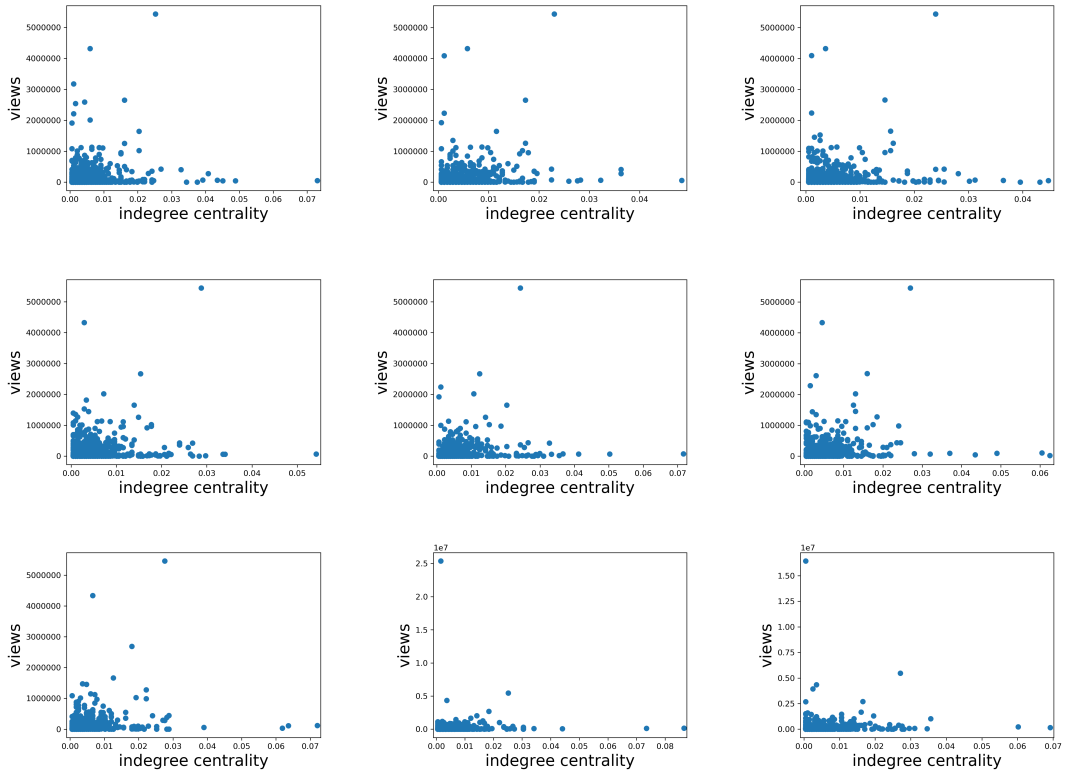


FIGURE F.5: Number of views versus indegree centralities. Plots refer videos collected on days 0,1,2,3,4,5,6,7 and 10.

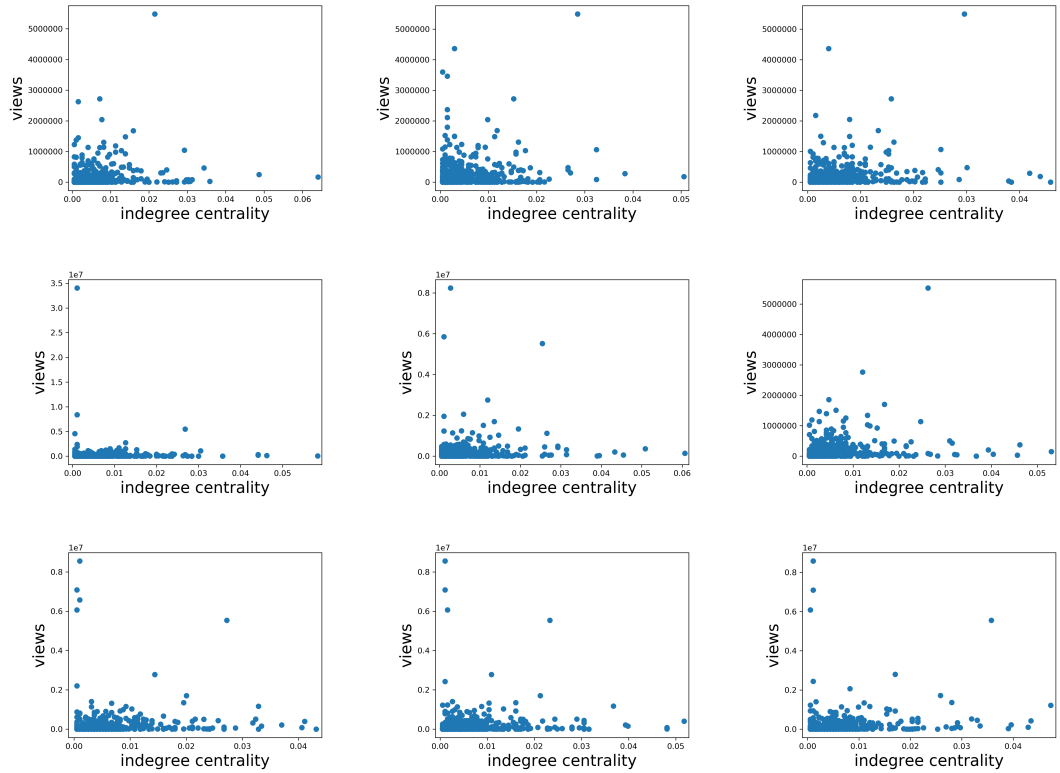


FIGURE F.6: Number of views versus indegree centralities. Plots refer videos collected on days 11,12,13,14,16,17,18,19 and 20.

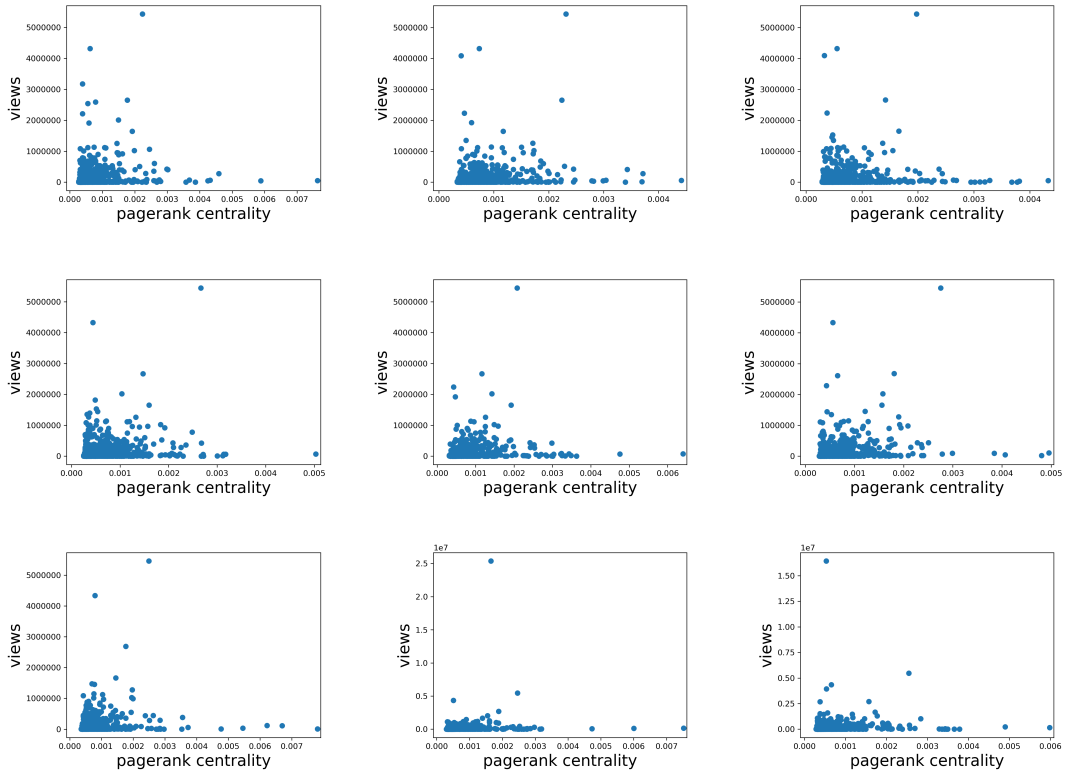


FIGURE F.7: Number of views versus PageRank centralities. Plots refer videos collected on days 0,1,2,3,4,5,6,7 and 10.

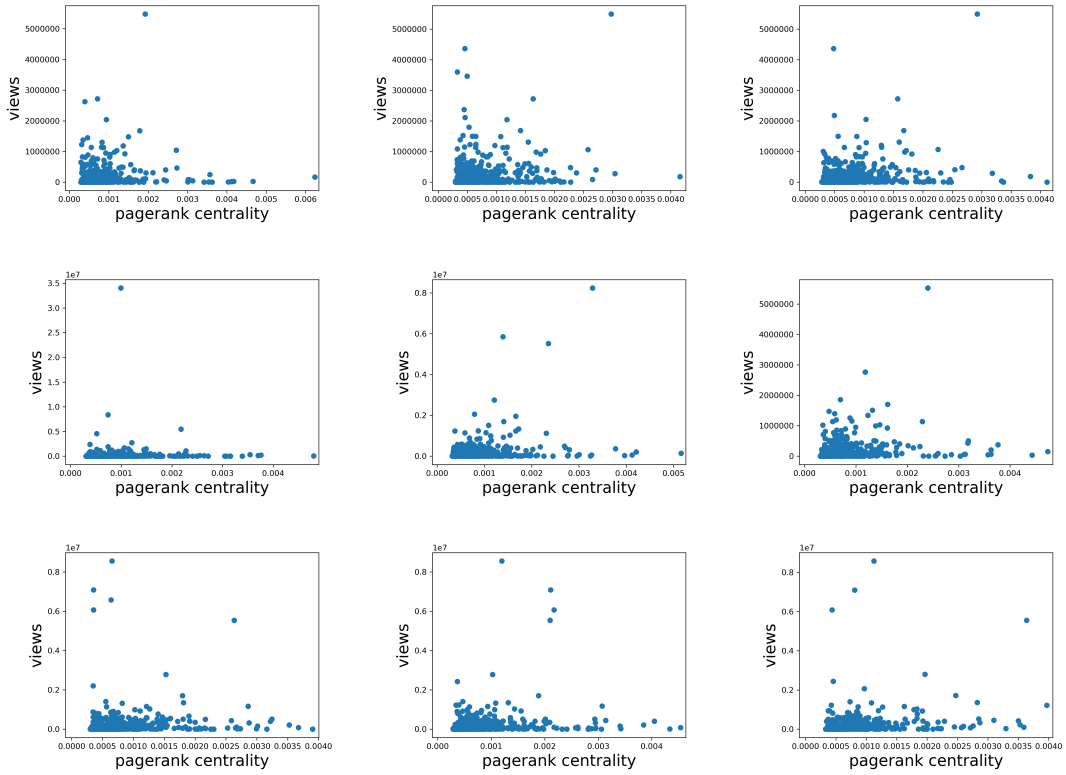


FIGURE F.8: Number of views versus PageRank centralities. Plots refer videos collected on days 11,12,13,14,16,17,18,19 and 20.

BIBLIOGRAPHY

[1] *2018 Italian General Elections*. http://www.interno.gov.it/it/speciali/2018-elections. 2018.

[2] M. Almeida-Neto et al. "A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement". In: *Oikos* 117.8 (2008), pp. 1227–1239.

[3] S. Aral, L. Muchnik, and A. Sundararajan. "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks". In: *Proceedings of the National Academy of Sciences* 106.51 (2009), pp. 21544–21549.

[4] S. Asur et al. "Trends in social media: Persistence and decay". In: *Fifth international AAAI conference on weblogs and social media*. 2011.

[5] R. Axelrod. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization". In: *Journal of Conflict Resolution* 41.2 (1997), pp. 203–226. DOI: 10.1177/0022002797041002001.

[6] S. Azaele et al. "Statistical mechanics of ecological systems: Neutral theory and beyond". In: *Reviews of Modern Physics* 88.3 (2016), p. 035003.

[7] P. Bak. *How nature works: the science of self-organized criticality*. Springer Science & Business Media, 2013.

[8] E. Bakshy et al. "The Role of Social Networks in Information Diffusion". In: *Proceedings of the 21st International Conference on World Wide Web*. WWW '12. Lyon, France: ACM, 2012, pp. 519–528. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187907. URL: http://doi.acm.org/10.1145/2187836.2187907.

[9] M. J. Barber. "Modularity and community detection in bipartite networks". In: *Physical Review E* 76.6 (2007), p. 066102.

[10] J. Bascompte et al. "The nested assembly of plant–animal mutualistic networks". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9383–9387.

[11] U. Bastolla et al. "The architecture of mutualistic networks minimizes competition and increases biodiversity". In: *Nature* 458.7241 (2009), p. 1018.

[12] C. Bender and S. Orszag. *Advanced mathematical methods for scientists and engineers*. International series in pure and applied mathematics. McGraw-Hill, 1978. ISBN: 9780070044524. URL: https://books.google.it/books?id=rwZRAAAAMAAJ.

[13] R. A. Bentley, M. W. Hahn, and S. J. Shennan. "Random drift and culture change". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271.1547 (2004), pp. 1443–1450.

[14] R. A. Bentley, P. Ormerod, and M. Batty. "Evolving social influence in large populations". In: *Behavioral ecology and sociobiology* 65.3 (2011), pp. 537–546.

[15] J. Borge-Holthoefer et al. "Emergence of Influential Spreaders in Modified Rumor Models". In: *Journal of Statistical Physics* 151.1 (2013), pp. 383–393. ISSN: 1572-9613. DOI: 10.1007/s10955-012-0595-6. URL: https://doi.org/10.1007/s10955-012-0595-6.

[16] S. Brin and L. Page. "The anatomy of a large-scale hypertextual web search engine". In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[17] J. W. Brown, R. V. Churchill, et al. *Complex variables and applications*. Boston: McGraw-Hill Higher Education, 2009.

[18] J. K. CAVERS. "On the Fast Fourier Transform Inversion of Probability Generating Functions". In: *IMA Journal of Applied Mathematics* 22.3 (Nov. 1978), pp. 275–282. ISSN: 0272-4960. DOI: 10.1093/imamat/22.3.275. eprint: http://oup.prod.sis.lan/imamat/article-pdf/22/3/275/2652991/22-3-275.pdf. URL: https://doi.org/10.1093/imamat/22.3.275.

[19] J. S. Clark. "Beyond neutral science". In: *Trends in Ecology & Evolution* 24.1 (2009), pp. 8–15.

[20] P. Covington, J. Adams, and E. Sargin. "Deep neural networks for youtube recommendations". In: *Proceedings of the 10th ACM conference on recommender systems*. ACM. 2016, pp. 191–198.

[21] C. De Bacco, D. B. Larremore, and C. Moore. "A physical model for efficient ranking in networks". In: *Science Advances* 4.7 (2018), eaar8260. ISSN: 2375-2548. DOI: 10.1126/sciadv.aar8260. URL: http://dx.doi.org/10.1126/sciadv.aar8260.

[22] M. Del Vicario et al. "The spreading of misinformation online". In: *Proceedings of the National Academy of Sciences* 113.3 (2016), pp. 554–559.

[23] A. Della Vecchia and C. De Bacco. "A physically-inspired model for dynamical ranking in networks". In: *Under preparation* (2019).

[24] T. Evans and A. Plato. "Exact solution for the time evolution of network rewiring models". In: *Physical Review E* 75.5 (2007), p. 056101.

[25] S. Fortunato. "Community detection in graphs". In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[26] C. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer complexity. Springer, 2004. ISBN: 9783540208822. URL: https://books.google.it/books?id=wLm7QgAACAAJ.

[27] M. Gell-Mann. "What is complexity? Remarks on simplicity and complexity by the Nobel Prize-winning author of The Quark and the Jaguar". In: *Complexity* 1.1 (1995), pp. 16–19. DOI: 10.1002/cplx.6130010105. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.6130010105. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cplx.6130010105.

[28] J. P. Gleeson et al. *Competition-Induced Criticality in a Model of Meme Popularity*. Vol. 112. American Physical Society, 2014, p. 048701. DOI: 10.1103/PhysRevLett.112.048701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.112.048701.

[29] L. K. Hansen et al. "Good friends, bad news-affect and virality in twitter". In: *Future information technology*. Springer, 2011, pp. 34–43.

[30] P. N. Howard et al. "Opening closed regimes: what was the role of social media during the Arab Spring?" In: *Available at SSRN 2595096* (2011).

[31] S. Hubbel. "The Unified Theory of Biodiversity and Biogeography". In: *Princeton Univeristy Press* (2001).

[32] A. L. Hughes and L. Palen. "Twitter adoption and use in mass convergence and emergency events". In: *International journal of emergency management* 6.3-4 (2009), pp. 248–260.

[33] J. P. Keener. *Principles of Applied Mathematics: Transformation and Approximation*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1988. ISBN: 0-201-15674-1.

[34] J. K. Lee et al. "Social media, network heterogeneity, and opinion polarization". In: *Journal of communication* 64.4 (2014), pp. 702–722.

[35] K. Lerman and R. Ghosh. "Information contagion: An empirical study of the spread of news on digg and twitter social networks". In: *Fourth International AAAI Conference on Weblogs and Social Media*. 2010.

[36] P. Lewis and E. McCormick. *How an ex-YouTube insider investigated its secret algorithm*. 2018. URL: https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot.

[37] M. Marder. "Dynamics of epidemics on random networks". In: *Phys. Rev. E* 75 (6 2007), p. 066103. DOI: 10.1103/PhysRevE.75.066103. URL: https://link.aps.org/doi/10.1103/PhysRevE.75.066103.

[38] R. M. May. "Ecological science and tomorrow's world". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1537 (2010), pp. 41–47.

[39] J. Moody. "Race, School Integration, and Friendship Segregation in America". In: *American Journal of Sociology* 107 (Nov. 2001), pp. 679–716. DOI: 10.1086/ajs.2001.107.issue-3.

[40] S. Nee. "The neutral theory of biodiversity: do the numbers add up?" In: *Functional Ecology* 19.1 (2005), pp. 173–176.

[41] M. Newman. *Networks: An Introduction*. OUP Oxford, 2010. ISBN: 9780199206650. URL: https://books.google.it/books?id=q7HVtpYVfC0C.

[42] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. "Random graphs with arbitrary degree distributions and their applications". In: *Phys. Rev. E* 64 (2 2001), p. 026118. DOI: 10.1103/PhysRevE.64.026118. URL: https://link.aps.org/doi/10.1103/PhysRevE.64.026118.

[43] C. Olston, M. Najork, et al. "Web crawling". In: *Foundations and Trends® in Information Retrieval* 4.3 (2010), pp. 175–246.

[44] N. Rahman et al. "A course in theoretical statistics: For sixth forms". In: *Technical Colleges, Colleges of Education, Universities: Charles Griffin & Company Limited* (1968).

[45] R. E. Ricklefs and S. S. Renner. "Global correlations in tropical tree species richness and abundance reject neutrality". In: *Science* 335.6067 (2012), pp. 464–467.

[46] D. M. Romero, B. Meeder, and J. Kleinberg. "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter". In: *Proceedings of the 20th international conference on World wide web*. ACM. 2011, pp. 695–704.

[47] D. M. Romero et al. "Influence and Passivity in Social Media (August 4, 2010)". In: *Available at SSRN 1653135* (2011).

[48] T. C. Schelling. "Dynamic models of segregation". In: *The Journal of Mathematical Sociology* 1.2 (1971), pp. 143–186. DOI: 10.1080/0022250X.1971.9989794. eprint: https://doi.org/10.1080/0022250X.1971.9989794. URL: https://doi.org/10.1080/0022250X.1971.9989794.

[49] A. Solé-Ribalta et al. "Disentangling co-occurrence patterns in n-partite ecosystems". In: *arXiv preprint arXiv:1807.04666* (2018).

[50] S. Suweis et al. "Emergence of structural and dynamical properties of ecological mutualistic networks". In: *Nature* 500.7463 (2013), p. 449.

[51] Twitter.com. *Twitter turns six*. 2012. URL: https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html.

[52] USAToday. *Twitter overcounted active users since 2014, shares surge on profit hopes*. 2017. URL: https://eu.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/.

[53] I. Volkov et al. "Density dependence explains tree species abundance and diversity in tropical forests". In: *Nature* 438.7068 (2005), p. 658.

[54] I. Volkov et al. "Neutral theory and relative species abundance in ecology". In: *Nature* 424.6952 (2003), p. 1035.

[55] I. Volkov et al. "Patterns of relative species abundance in rainforests and coral reefs". In: *Nature* 450.7166 (2007), pp. 45–49. DOI: 10.1038/nature06197. URL: https://doi.org/10.1038/nature06197.

[56] D. J. Watts. "A simple model of global cascades on random networks". In: *Proceedings of the National Academy of Sciences* 99.9 (2002), pp. 5766–5771.

[57] L. Weng et al. “Competition among memes in a world with limited attention”. In: *Scientific Reports* 2 (Mar. 2012), 335 EP –. URL: https://doi.org/10.1038/srep00335.

[58] H. S. Wilf. *Generatingfunctionology*. Natick, MA, USA: A. K. Peters, Ltd., 2006. ISBN: 1568812795.

[59] F. Wu and B. Huberman. “A persistence paradox”. In: *First Monday* 15.1 (2010).

[60] S. Zapperi, K. B. Lauritsen, and H. E. Stanley. “Self-organized branching processes: mean-field theory for avalanches”. In: *Physical review letters* 75.22 (1995), p. 4071.