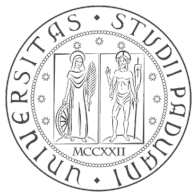

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale
in Ingegneria delle Telecomunicazioni

Epidemic models for research ideas spreading in the scientific community

Laureando

Anna Valeria Guglielmi

Relatore

Leonardo Badia

Anno Accademico 2013/2014

Contents

Abstract

In this thesis, we apply the basic epidemiological *SIR model* and some of its extended versions, *SEIR model* and *SIR model* with demography, to the propagation of scientific ideas in the worldwide community of researchers, in order to investigate the spread of those ideas. To this end, we collected quantitative records of articles published in scientific conferences for some selected topics, such as big data, software defined networking, LTE advanced, cloud computing, Internet of things, game theory, Bluetooth, and DVB-T, over a 5-year period with a monthly granularity in order to better verify the applicability of different epidemic models. The values of the *basic reproductive ratio*, which indicates the maximum reproductive potential for an infection, are discussed to compile a classification based on the contagion level of these topics, and the types of description that the different models give are investigated to discuss according to their approximation level and their descriptive potential.

In questa tesi, applichiamo il modello epidemiologico *SIR* di base e alcune delle sue versioni estese, *modello SEIR* e *modello SIR* con demografia, alla diffusione delle idee scientifiche nella comunità globale dei ricercatori, in modo da investigare sulla diffusione di tali idee. A tale scopo abbiamo raccolto il numero di articoli relativi ad alcuni argomenti pubblicati in conferenze scientifiche. Gli argomenti scelti sono big data, software defined networking, LTE advanced, cloud computing, Internet of things, Bluetooth e DVB-T, e l'intervallo temporale considerato è di 5 anni in modo da poter verificare meglio l'appropriatezza dei diversi modelli epidemici. Abbiamo discusso i valori ottenuti per il *basic reproductive ratio*, che indica il massimo potenziale riproduttivo di un'infezione, in modo da stilare una classifica basata sul livello di contagio degli argomenti. Infine, abbiamo analizzato le tipologie di descrizioni date dai diversi modelli per discutere sul loro livello di approssimazione e sul loro potenziale descrittivo.

Chapter 1

Introduction

Since the early decades of the twentieth century, the evaluation of scientific developments through some quantitative tools of measurement has caught interest. Initially, the purpose was to give prestige and importance to a specific scientific field rather than others [1;2;3]; thus, the first statistical investigations appeared, based on the literature and previous works. Then, in the second half of the 1900s, many people started to focus on more general points of view, and in this way a complete evaluation of science in all its forms was made possible [11;12]. These evaluations were increasingly based on manual mechanisms, such as counting the papers that a scientist publishes on a specific topic, or the citations that a given article received, and so on; such methods allowed the definition and the development of a statistical analysis, bibliometrics, which was deeply different from other forms of investigation applied to the records of human knowledge [13;14;15;17;18]. Indeed, bibliometrics is a discipline with applications in the history and sociology of knowledge, in clerical activity of library archives, as well as in technical fields such as communication and information science; it revolves around the measurable properties of the systems and technologies that vehicaulate knowledge. Moreover, bibliometrics establishes connections between documents as scientists expose their findings and research to their peer community.

It is reasonable that in the last few years, with the advent of technologies, such as the Internet, the basic bibliometric methods have been developed also

in these elaborate systems. Web of Science [26] and Google Scholar [28] are two well-known examples of online databases based on the last innovative bibliometric indicators that allow the user to search and compare on a large scale most of the topics in the scientific literature. Through the investigations that these tools allow, it is possible to design a complete map of science, in which all the scientific documents present are connected through some references or citations to the others.

The network of citations has been characterized extensively by researchers investigating complex networks [81;82]. The bibliometric mechanism of counting the number of publications related to specific selected topics has been used to find empirical data essential to the evaluation of idea propagation, but this is not the only tool. In this thesis, the objective is to investigate an efficient model to evaluate the spreading of research ideas in the scientific community. We focus not on the structure of the citation system related to an idea, but on the idea itself.

To this end, we argue that the propagation of a research ideas could be compared with the spread of an infectious disease. Similar to a pathogen responsible for an infection, a scientific idea could be considered the agent that caused the spread of a certain topic in the scientific community. In epidemiology, many models have been developed to understand the dynamics of infectious diseases [30;31;34;33;38;39;40], the most popular and important being the *SIR model* conceived by Kermack and McKendrick and its successive variations and improvements [32;44]. Through it, other models much more complex and disease-specific have been developed to study and evaluate the main causes of human mortality due to infections [41;42;46;47;48].

In such a system, the relationship between the infectious disease spread and the more general concept of network develops. Instruments for network analysis and structured characterization have also been used to improve the realism of the models for infection propagation [53;54;56;57]. On the other hand, it is possible to find in the literature some applications of the epidemiological models to solve relevant network problems in computer science and telecommunications. For example, the propagation of viruses in Mobile Ad Hoc Networks (MANETs) has been compared to the spread of an infection

in a population, thus the basic *SIR model* has been employed to analyze it [65]. Some routing protocols have been designed inspired by the process of contacts between susceptibles and infectives and the related transfer of an infection [58]. Other examples of applications of epidemic models to general networks are [59;61;63;64].

People commonly refer to "infectious" slogans, or "viral" marketing campaigns, or ideas that "spread like wildfire." These everyday phrases indicate a basic affinity between the propagation of ideas and the transmission of infectious diseases; indeed, both are processes in which something is communicated, that is, transferred from one person to another. In 1964, the first scientist that highlighted the relationship between epidemics spread and the diffusion of ideas was W. Goffman [70]; after explaining the basic assumptions of the well-known epidemiological model, he tried to generalize the epidemic theory in order to apply it to the transmission of ideas. His initial purpose was to determine the perfect circumstances under which it may be necessary to introduce an information retrieval system to help scientists in their research. Later, having thoroughly studied the epidemiological models, Goffman focused on the spread of knowledge about mast cells [71]; he applied the *SIR model* considering the population as the list of authors who had written on mast cells. This list was based on a bibliography created earlier by Hans Selye. At the time of Goffman these type of studies were time-consuming; it was difficult to find experimental data, that is the data that could represent the population in the epidemiological model, because it involved a totally manual search. In general, science evaluation methods were not well established; in particular, there was talk about scientometrics, that is the study of measuring and analysing science, technology, and innovation throughout measurement of their impact. At that time the most likely candidates were scientific citations and the definition of representative articles to investigate the impact of journals and periodicals. Thus, the application of epidemic models was just a tentative proposal.

Nowadays, the development of online databases that keep track of all the publications of papers, articles, journals and books, allows a faster research about specific topics. Recently, only two other works, other than that de-

veloped by Goffman, have been identified [72;73]. In [72] the authors apply several paradigmatic epidemic models to empirical data on the advent and propagation of Feynman diagrams through the theoretical physics community in the USA, URSS, and Japan in the period immediately after World War 2; instead, in [73] a new approach, based on epidemic models modified in individual-based weighted network models, is used to describe the spread of research topics across different disciplines. In this example, the empirical data were obtained throughout the use and the consideration of citations between several articles.

In this thesis, we will apply the basic epidemiological *SIR model* and some of its extended versions to the counts of the number of articles published for several scientific topics in order to investigate the spread of those ideas in the scientific community. We will discuss the related values of the *basic reproductive ratio*, which indicates the maximum reproductive potential for an infection, and the types of description that the different models give. In general, there is a good qualitative agreement at the descriptions. In some cases the epidemic model is very accurate and also able to predict future developments; in the remaining cases, the model can be improved to better approximate the experimental dynamics.

The rest of this thesis is organized as follows: in chapter 2, the main concepts of bibliometrics and epidemic models are discussed with their applications. In chapter 3, the models and the assumptions used to evaluate the research ideas spreading in the scientific community are described. Chapter 4 discusses the main results. Finally, some concluding remarks are presented in chapter 5.

Chapter 2

Related work

2.1 Bibliometrics

2.1.1 History and Definition

The first systematic collection of statistics on science is attributed to the American psychologist James McKeen Cattell, that, in the 1906 launched the biographical directory *American Men of Science*, published periodically, wherewith he collected information on the scientists active in the United States, [76]. From these data, Cattell conducted systematic and regular studies on science and its development until 1930s; therefore, he produced some measurements and statistics on the number of scientists and their geographical distribution, and categorized scientists on the basis of the performance evaluated. For these reasons, Cattell can be credited for having started the mechanisms of systematic measurements, that has allowed the development and diffusion of *bibliometrics*.

Cattell introduced two kinds of parameters in the measurement of science: *quantity* and *quality*. The former, also called *productivity*, was simply counting the number of scientists a nation produce. The latter, also called *performance*, and measured by averaging peer rankings of colleagues, was seen as the contribution to the advancement of science. The first use of the directory in the statistical analysis was concerned with psychologists: in the 1930s Cattell analyzed the "academic origin [institution], course and destina-

tion" for a selected group of 200 psychologists [1]. The purpose of the study was to identify the best scientists, display their performance, compare different nations and suggest some future actions in the field. Another important goal of the application of the systematic collection was proving the influence and the prestige of psychology and that it was really a science among the sciences [2].

B. G. Miner, another psychologist, presented a more systematic analysis to further support this point. In [3], he said that "in the following pages the writer collects certain facts which bear upon the recent development of psychology in American institutions of higher learning, with the hope of giving a more adeguate means for judging the present status of science." The data used by Miner were taken from a list of 150 colleges and given by the directors of 34 prestigious laboratories. Miner presented statistics on these laboratories and classified them according to equipment and apparatus. He also presented numbers of income and quantified the share of effort dedicated to psychology in universities; he also considered chairs, departments and their sub-divisions or specialties, professors and courses and he made some assessments about the percentage of university enrollment in psychology. Furthermore, Miner tried to categorize the universities and the courses according to their number of doctorates.

Cattell and Miner inspired other psychologists with their use of statistics to proving the importance of psychology. Some other examples are E. F. Bucher [4;5], C. A. Ruckmich [6;7], S. W. Fernberger and S. I. Franz [8;9]. They published several periodic reviews, some of these were strictly qualitative, but others included quantitative materials according to the numbers given by several performance metrics evaluated. Bucher, for example, reviewed the work of the American Psychology Association, the largest scientific and professional organization of psychologists in the United States, and its influence, in terms of laboratories and systematic literature of its members, and classified and calculated the number of papers presented over the decade by psychologists. To Bucher, publication and paper counts provided "a good measurement of the annual variation of the intensity of interest in the generic topics with which the psychologists are engaged" [5]. Later, in

1912, another psychologist, C. A. Ruckmich of Cornell University, published a review of 25 years of psychology that was all based on statistics [6]: in this work, he measured laboratories, courses, department, and their conditions. Furthermore, he compared psychology to other disciplines in terms of number of professors, academic hours, registrations, and appropriations. Another review was published by Ruckmich in 1916 [7]: he selected six journals from 1905 to 1915 and counted the number of papers, the number of pages dedicated to each article, and classified their methodology. The statistics on publications was further developed by S. W. Fernberger; in [8], he evaluated the evolution of membership to the American Psychological Association, and the increasing emphasis placed on publishing as a criterion for suitability, and discussed the finances, journals, organization, and meetings of the Association for the evaluation of the research interest.

In this same period, Fernberger and Franz introduced the idea that the only count of the publications or papers was not sufficient to evaluate the performance and the statistics. It was necessary to consider the entire background; indeed, even the age difference of the authors influences the statistics of discipline productivity: Franz looked at the date of the doctorates as the date when publications might reasonably be expected and compared the number of actual contributors versus the expected ones; he found that the actual contributors in relation to the expected ones decreased. This is due the fact that some of the contributors according to the age may not necessarily be active over the whole period considered for the statistic evaluation. Furthermore, he measured that the productivity of the older researcher, where old/young depended on the date in which they were granted their doctorate, was higher than the younger ones, but the ratio of actual to expected publications was higher among the younger scientistis. Another important aspect was the author's nationality: to evaluate the development of the science through systematic collections and to have global view and comparison, it was necessary to consider also the publications of the other nations, not only the works that came from the American Psychological Association. Fernberger started a series of papers on the international comparison of scientific production on psychology [9]. Because these papers were published

periodically every ten years and from the analysis of the results collected, Fernberger could evaluate the so called "political economy" of the research [10]. He looked at the effects of the world wars, politics, and nationalism on the measurements of the publications; for example, a result that he found is that war and subsequent economic crisis tend to decrease the number of publications. On the other hand, the presence of a strong government that encourages and even supports research with funding considerably increases the amount of scientific publications.

Still, publication counts were limited to one subject, and not to the whole community of scientists. The first to perform this leap was W. Dennis who looked at the most well-known and celebrated scientists and their scientific production. In his paper [11], he chose 41 scientists whose names appeared between 1943 and 1953 in the *Biographical Memoirs* of the US National Academy of Science and who reached the age of 70, from their biographies he calculated the number of publications of each and found that these scientists have been globally responsible for 203 papers per year on average. Later Dennis looked at the 25 most distinguished scientists of the nineteenth century, this evaluation was based to the space devoted to them in encyclopedias and dictionaries of biography: he said that in "science, quantity and quality are correlated." In another publication, [12], Dennis analyzed the age at which scientists produce most by counting the number of papers of a selected group of them.

Simultaneously to the several investigations led by these psychologists, in the first half of the twentieth century other sporadic works in other different disciplines used statistical analysis of publication: in 1917, Francis J. Cole and Nellie B. Eales applied quantitative analysis to the comparative anatomy literature from 1543 through 1860 [13]. They represented with a curve the documentary growth rate over the period considered, they tried to determine the aspects of this subject that had most attracted scholars efforts in time and correlate evolution and recession phases of research activity with economic, social and human factors. This was credited as the first real bibliometric investigation, together with the previously cited publications of Fernenberg, mainly related to psychology, and Dennis, more general in scope; indeed,

they analysed science and scientific progress through statistical analysis of publications, they used the results as a tool for mapping scientific research and locating the most eminent scientists, and they considered the limits of such a method because numbers alone did not tell the whole story of science; as a consequence, the quantitative analysis must be associated with a qualitative analysis of scientific literature values.

The study of Cole and Eales was followed by a statistical analysis of the history of science in the 1923 by E. Windham Hulme, a librarian of the British Patent Office [14]. The key idea of Hulme was that by classifying all the books in the world according to some universal criteria and ordering them chronologically within each topic, a map of the human mind could be worked out. So, scientific specialization reflected the process of human civilization. His analysis in [14] was based on the journal entries in the seventeen sections of the *International Catalogue of Scientific Literature* and he presented the rankings of entries in physiology, bacteriology, serology, and other medical subdisciplines, the rankings of sciences according to their output of periodical literature, the number of journals in the annual issues arranged by subject, and the number of indexed journals arranged by country of publication. In this scenario, the idea that there was a relationship between the macrocosm of human knowledge and the microcosm of the library and the collection of books and journals had to deal with the practical constraints imposed by restricted budgets and the lack of physical space against an increasing volume of potential relevant documentation. According to this intuition, a quantitative insight into bibliographies, library collections, and catalogs offered several advantages. In 1927, P. Gross and E. M. Gross claimed that it was no longer sufficient to "sit down and compile a list of those journals which one considers indispensable" because "often the result would be seasoned too much by the needs, likes and dislikes of the compiler." In [15], they counted and analyzed the citations in articles published in a chemistry journal, the prestigious *Journal of the American Chemical Society*, and they wrote a list of journals they considered indispensable in chemical education by ranking the journal title according to the number of citations received. Furthermore, they claimed that the citation count was not the only reasonable criterion to

consider for journal selection, because the age distribution of the references produced a similarly important index of utility; indeed, according to their idea, if some journals had the same number of citations, the journals receving citations to the most recently published articles were classified higher because the present trend was more significant than the past performance of a journal. In this way, they established a relationship between quality, citation rates, and time distribution of citations; as a result, they revolutionized the analysis of the information for library management and research evaluation purposes. This was the first study based on counting and evaluating citations, an idea that inspired Eugene Garfield later for the creation of the *Science Citation Index*. It should be noted that the previous studies were based on entries in bibliographies, not on received citations.

A further turning point was represented by Paul Otlet, father of European documentation and cofounder of the International Institute of Bibliography in the late 1920s. He clearly distinguished what Pritchard called later *bibliometrics* from other forms of statistical investigation applied to the records of human knowledge. In [16], Otlet celebrated the measure as a superior form of knowledge and supported the development of a subfield of bibliology entirely dedicated to the collection of measures related to documents and papers of all kinds.

But it was Alan Pritchard that in the 1969 coined the term *bibliometrics* [17]. He defined bibliometrics as the application of mathematics and statistical methods to shed light on the processes of communication, the nature and the development of a discipline, by means of counting and evaluating the several aspects of written comunication. In later articles, Pritchard explained bibliometrics as the "metrology of the information transfer process and its purpose was analysis and control of the process" [18]. Pritchard's interpretation was upon the fact that measurement is "the common theme through definitions and purposes of bibliometrics and the things that we are measuring when we carry out a bibliometric study are the process variable in the information transfer process" [18]. Then other formal definitions have been attributed to bibliometrics; for example, the *British Standard Glossary of Documentation of Terms* defined bibliometrics similarly to Pritchard. An-

other example of definition is given by William Gray Potter, editor of the issue of *Library Trends* dedicated to bibliometrics: "bibliometrics is, simply put, the study and measurement of the publication patterns of all forms of written communication and their authors" [19]. In the same issue, Alvin M. Schrader said "bibliometrics is the scientific study of recorded discourse." All these, and other, subsequent definitions clarified and deepened the purposes and the field of study (number of publications per authors, type of publications, etc.) of bibliometrics, expanding its use to the quantitative analysis of scientific productivity on a large scale.

A new historical phase of bibliometrics began with the end of World War 2, when a new political, social, and economical arrangement, and a new organization of the human scientific knowledge was established in the background. In this scenario, much more complex in terms of scientific productivity, all the tools and the ideas of measuring the scientific documentation of the first statistic collections were recovered. Furthermore, there was certainty that scientific activity could be controlled, planned, and addressed towards important purposes, mainly because there was a strict relationship between the economical growth and the development and innovation carried by science. Later in the politics of the scientific research, a significative change came from the hurl of *Sputnik* by the Soviet Union in the 1957: a riorganization of the American research system was necessary to fill the scientific gap towards the Soviets. This new organization, however, was interested more in the human and finance resources rather than in the results of the research itself, while the evaluation of the research still belonged to the internal mechanisms of the academic communication. The attention to the measurement of the scientific documentation was addressed not only for the evaluation of the research; in that same period (1950-1960), indeed, there was a notable increasing of the scientific literature that presented the need to search new control instruments. The amount of the scientific production became a serious problem for those who needed to identify and make use of much more relevant information for their research, furthermore significant limits were due to the manual sistems of literature indexing.

At this point of the bibliometrics history the figure of Eugene Garfield

becomes meaningful. Garfield realized the inadequacy of the tools that scientists used to recover useful information according to their activity of research. He presented in the journal *Science* (1955) the idea of a project of interdisciplinary index based on citations, but only in 1963 he published the *Science Citation Index*; in [20] he said: "In this paper I propose a bibliographic system for science literature that can eliminate the uncritical citation of fraudulent, incomplete, or obsolete data by making it possible for the conscientious scholar to be aware of criticisms of earlier papers. Even if there were no other use for a citation index than that of minimizing the citation of poor data, the index would be well worth the effort required to compile it." His target was to devise a citation index that could offer a new approach to subject control of the literature of science, he talked about the *Science Citation Index* as "an association-of-ideas index." The term *impact factor* began to be used; it was similar to the quantitative measure obtained by Gross in evaluating the importance of scientific journals. Furthermore, Garfield emphasized that this factor was much more indicative than an absolute count of the number of publications of a scientist, as used for example by Dennis.

The *SCI* designed by Garfield further developed the correspondence between the qualitative value of periodicals and the number of citations received, stating the use of citations as evaluative tool and, indirectly, of the bibliometric indicators in the evaluation of the scientific research, that, since the 1960-70s, asserted themselves first in the United States and then in several European countries. Several studies based on the data derived from SCI were published; the first example is *Science Indicators Reports* published in 1973, edited by *National Science Board*, the purpose of which was the measurement of the American scientific research. Another important study focused on Garfield's idea was that of Narin, the *Evaluative bibliometrics* [20], in which he advanced the concept of citation impact as a qualitative measure of the scientific publications at the international level. That work was approved by the *National Science Foundation*, a United States government agency that supports fundamental research and education in all the fields of science and engineering.

Governments and institutions too were interested in the improvement of

the bibliometrics tools as a valid support for the decisions which provided for the distribution of the resources to invest in research. Afterwards, while in 1970s the American studies were focused mainly on the organizzation of the scientific literature, in Europe in the 1980s the bibliometrics was used to evaluate the performance of the universities, institutions, and research groups. It is important to highlight that in the last decade many innovative aspects were introduced in the field of bibliometrics: the creation of new open access citation indexes that have determinated the end of the de facto monopoly dictated by *SCI*, the development of much more sophisticated new bibliometrics indicators, and the application of bibliometrics in increasingly wider areas, like, for example, the World Wide Web.

2.1.2 Theoretical Basis and Bibliometrics of our days

A real watershed in this discipline is represented by the studies conducted by A. Lotka about the distribution of the scientific production in relation with the number of authors, by G. K. Zipf about the distribution of the words in documents, and by S. C. Bradford about the distribution of the number of articles in periodicals. These were the results of a quantitative methodological approach arised in the field of information and scientific communication in the first half of the twentieth century. These mathematical foundations of bibliometrics are not comparable to the way Newton's laws of motion and gravitation are the foundation of classical mechanics; indeed, the laws of bibliometrics do not allow to predict the number of articles an author will write, the number of citations that a paper or a publication will receive over a certain time span, or the number of journals that will publish papers on a given topic, but they allow to meaningfully combine the structure of several existing databases.

In the 1926 Alfred Lotka, mathematician and chemist, president of the *American Statistical Society*, published a study of the distribution of the scientific production in communities of scientists, chemists, and physicists [22]. He was interested in determining "the part which men of different calibrate contribute to the progress of science." From this study, indeed, Lotka ob-

served that the scientists contributed in a different way to the improvement of the knowledge: there was a small number of authors with high rate of productivity against a much more greater number with low output. He summarized these results in an empirical law according to which the number of authors that produce n publications is approximately equal to $1/n^2$ of the number of authors that publish a single article, and the ratio of all authors that produce a single article is approximately equal to the 60% of the entire production. Therefore, the value of Lotka's law was strictly related to the average productivity of a specific scientific community, that depended on the discipline evaluated. In general, Lotka's law is close to the observed values when applied to a discipline in which the level of authors productivity is very low.

After 8 years since the publication of Lotka's law, another empirical law was hypothesized by Samuel C. Bradford, librarian at the science museum in London, who published the results of his study in [23]. Bradford said that "the aggregate number of articles in a given subject, apart from those produced by the first group of large producers (periodicals), is proportional to the logarithm of the number of producers concerned, when these are arranged in order of decreasing productivity." In other words, this means that if periodicals contributing to a subject are ranked and then grouped in such a way that each group contributes the same number of articles, the numbers of periodicals in each group increase geometrically. Therefore, it was possible to identify a core of publications considered essential within each scientific discipline, and organize and manage bibliographic collections in a better way: not all the publications were necessary, or at least the benefits derived from them would have been useless in relation with the costs that would have been incurred. Furthermore, by optimizing the selection of material around the core of relevant publications, also the bibliographic research obtained benefits with the increase of the precision and the consequent noise decrease. Bradford's law was used by Garfield to choose the main periodicals that arranged the core of the *Science Citation Index*.

The third major study was to George K. Zipf, an American linguist and philologist who formulated a law valid both for bibliometrics and quantitative

linguistics. Zipf's law was derived from the study of the regularity of words in a text. In [24], Zipf stated that if words are ranked according to their frequency of occurrence, the nth ranking word will appear approximately k/n times, with k a proper constant. In other tems, Zipf's law measured the meaningfulness of a word as a function of the frequency with which it appeared in periodicals. Therefore, it was also possible to define a core of words in a text that were more relevant because more frequent.

From the historical development of bibliometrics, it can be seen how from a simple count of publications related to specific disciplines or more generic areas, the quantitative and qualitative analysis of scientific literature has increased the use of more sofisticated tools based on the importance of *citations*. The use of citations as an index for the identification of prestige in the scientific community began to spread. Not all the representatives of the scientific community were in agreement, some argued that the count of citations did not reflect the effect of the scientific activity: the choice of the citations for a publication was not so much based on the content of the quoted text, but mainly on some characteristics as the importance of the author, the editorial position of the text or the importance that the scientific community gave to that text. For example, Van Raan in [67] observed that the number of citations also depended on a multiplicity of factors, such as time, indeed the probability of beeing quoted was higher as the date of publications of the quoted article was closer, or the fact that citation procedures varied in the different scientific community, or factors connected to some characteristics of the periodical in which the article was published, as the frequency of publication of such periodical or the order in which the article was inserted in a periodical and so on. Another example is Collins who in the study of the citations of a group of papers about physics found that the number of citations, but in particular the acceptance of new innovative ideas in the physical community, depended on several variables different from the content of the quoted articles, such as, for example, the institutional context, [69]. On the other hand, there were someone else that was inclined to the first idea; for example, Baldi, studying the citations in the field of astrophysics, concluded that the characteristics different from the content of the articles

did not influence significantly the probability of beeing inserted in the list of references [68].

Today there is a shared consensus the idea that citations are not an *ideal* measurement of the scientific performance, but they are a good indicator of the notoriety of a group of research rather than of a single contribution [25]. A procedure of evaluation that takes care only of the number of citations that an article receives, without considering the meaning of these citations or the quotation behavior, could induce some erroneous considerations on the final result. For that reason, normally we do not speak only of counting the number of citations of an article, but we talk about *co-citation analysis* and *bibliographic coupling analysis*. The former means that when two document are quoted by the same third document, this represents an association between them; moreover, if two documents are quoted simultaneously over and over again the greater will be their association; while, the latter represents the ratio of two or more documents that cite a third.

With the advancements of the technology, bibliometrics could also develop and originate tools in the Web. A respectable successor of Garfield's *Science Citation Index* is *Web of Science*, WoS [26], a bibliographic database of citations and multidisciplinary online from 2002, accessible with fee through Thomson-Reuters' platform *Web of Knowledge*. The database WoS allows to search through about 12 000 periodicals, 148 000 conference proceedings, and 28 000 books; WoS includes 7 indices of citations: *Conference Proceedings Citation Index*, *Science Citation Index Expanded*, *Social Sciences Citation Index*, *Arts & Humanities Citation Index*, *Index Chemicus*, *Current Chemical Reactions*, *Book Citation Index*.

Selection in WoS is based on impact evaluations and comprise open-access journals, spanning multiple academic disciplines. The coverage includes: hard sciences, social sciences, humanities, and arts, and goes across disciplines. The seven citation indices listed above contain references, which have been quoted by other articles. One may use them to launch quoted reference search, that is, identifying articles that cite an earlier or current publication. Citation databases can be searched by author, topic, source title, or location. Another tool integrated in the *Web of Knowledge* platform, but different from

WoS, is the *Journal Citation Report* (JCR) that is split into two editions: the first contains data of more than 5 900 journals in 171 disciplines, the second contains data of more than 1 700 journals in 55 disciplines [26].

However, Web of Science does not index all journals, and its coverage in some fields is less complete than others. Furthermore, in 2009 the total file count of the WoS was 46.1 million records, which included 727 549 189 cited references. This citation service on average indexes around 65 million items per year, and it is described as the largest accessible citation database.

Another example of online citation index is *Scopus* [27], a citation database with fee launched in 2004 in the scientific, technological, biomedical field and in the area of social sciences. It is accessible from Elsevier's platform *SciVerse*. Scopus indexes 18 500 peer-reviewed periodicals, 1 800 open-access periodicals, around 4 million of conference papers, 425 business publications, 350 books, 375 million of web pages through *Scirus*, 24.8 million of patent registrations. Totally, *Scopus* contains 46 million registrations, whereof 25 million with references after 1996 and 21 million from 1823 to 1996. *Scopus* allows to set up a research by author, topic, source title, or location; it allows to visualize directly the abstract or the full text of the article sought, and it allows to set up some alert to inform about a specific topic or the publications of a specific author.

A database that is certainly well-known is *Google Scholar* [28]; it is a freely accessible web search engine that indexes the full text of academic and scientific literature across an array of publishing formats and disciplines. *Google Scholar* was beta released in November 2004, its index includes most peer-reviewed online journals of Europe and America's largest scholarly publishers, and in addition scholarly books and other non-peer reviewed journals; it also indexes technical reports and freely published documents on the WWW. It is similar to the subscription-based tools, Elsevier's *Scopus* and Thomson's *Web of Science*; it puts together traditional method of research based on metadata, citation linking, many features of bibliometrics and the new possibility to identify and recover the digital full text of a large number of different documentations with free access and the references of articles and digital or paper books. *Google Scholar* supplies informations about the

type of material that becomes part of his archive, but it does not outline the method of selection of materials, the criteria of inclusion, the update times and the procedure of indexing. As a consequence, it is not possible to control and to evaluate the reliability and the coverage of the research. Compared with *Web of Science* and *Scopus*, *Google Scholar* achieves its indexing results much more quickly but also with lower accuracy, especially for searches that require deep investigations.

These tools are all based on the use of bibliometrics indicators. The simplest are the count of the number of publications and the count of the number of citations, but other much more complex bibliometric indicators, such as impact factor or the recent H-index [66], are also available. The latter has became a widely used bibliometric indicator; according to its definition a scientist has index h if h of his/her Np papers have at least h citations each, and the other $(Np - h)$ papers have no more than h citations each [66]. Its aim is to express in a single numerical value both the productivity, that is the number of publications, and the effect on the scientific community, that is the number of citations, referred to the individual researcher.

2.2 SIR model

2.2.1 Epidemic models

Although chronic diseases such as cancer and heart conditions receive more attention in developed regions, infectious diseases are the most important cause of mortality in developing countries. Also, even in developed regions the human immunodeficiency virus (HIV), which is the etiological agent for acquired AIDS, nowadays is an important sexually-trasmitted disease throughout the world. Other diseases such as tuberculosis are becoming a problem because drug-resistant strains have evolved. In the past some massive epidemics destroyed entire populations. The *Black Death* in the 14th century is just the most famous epidemic historically. Moving across the Atlantic Ocean, the first major epidemic in the USA was the *Yellow Fever* epidemic in Philadelphia (1793) in which about 5 000 people died out of a

population of around 50 000 [83;84]. Another well known epidemic was *The Plauge of Athens* (430-428 BC), described by Thucydides; importantly, there was no mention of person-to-person contagion, a remark that led to a better comprehension on how epidemics can propagate (not only directly, but also through a proxy) [85].

Understanding the transmission characteristics of infectious diseases in populations, regions, and countries can lead to better approaches to contain the transmission of these diseases. Mathematical models are useful in building and testing several theories and comparing, implementing, and evaluating detection, prevention, and theraphy: the progress of an epidemic across a population is highly amenable to mathematical modelling.

Pioneered by Daniel Bernoulli in 1760 [29], mathematical modeling has a well-known history in predicting and rationalizing the spread or control of infectious diseases in a population. Bernoulli, a French matematician and physicist, made the first study on the problem concerning the spread of human diseases. His studies focused on the spread of smallpox; notably, in the early years of the eighteenth century the descendents of Louis XIV of France were killed off by this diseases, so the problem was topical at the time. His work, presented to the French parliament whereof Bernoulli was a member, impressed the assembly but did not achieved the hoped results. Only in the 1798, Edward Jenner published his well known discoveries on the smallpox vaccine, whose validation he also based on Bernoulli's studies.

The current literature is rich with epidemic models, which have enhanced our understanding of outbreaks, epidemics, and pandemics of various pathogens. Particularly, the principles enunciated by Hamer in 1906 [30] and later extended by Ronald Ross in 1911 [31] and Kermack and McKendrick in 1927 [32], establish the true foundations of mathematical epidemiology today. Hamer in [30] formulated and analysed a discrete time model in his attempt to understand the recurrence of measles epidemics; his model may have been the first to assume that the incidence of the disease depends on the product of the densities of the susceptibles to the disease and the infectives. Ross [31] has developed a system of differential equations to represent, in certain circumstances, the course of events in a community that has become infected

with malaria. He did not give the general solution of the equations, but he resctricted his discussion to the final state of equilibrium to which they lead. On the other hand, in [32] Kermack and McKendrick formulated the first generic model to study the spread of infectious diseases and obtained the epidemic threshold result according to which the density of susceptibles must exceed a critical value in order for an epidemic outbreak to occur. The model envisaged a population partitioned in 3 classes: the class of *susceptibles* to infection, where no pathogen is present, just a low-level nonspecific immunity within the host; the class of *infectious*, consisting of individuals with high pathogen level and the potential to transmit the infection to other susceptible individuals; the class of *recovered*, which includes individuals naturally resistant to the pathogen or those who cleared the infection and are no longer contagious, nor they can be infected again. The model of Kermack and McKendrick provided that, given a population of individuals in which contagion is instantaneous and happens with rate C, if M is the death rate for the epidemic and k is the percentage of meeting between healthy and sick individuals, then the disease spreads or not depending on whether the initial number of susceptibles is larger than or smaller than kM/C, respectively. This is the first idea of *SIR model* that inspired many other epidemiological models.

Although it was well known that animal/human hosts and their parasites varied in resistence and infectivity respectively, and that many other factors played their part in how an epidemic disease spreads there were some epidemiologists, such as Greenwood [45], that expressed criticism on the productivity of the model. Greenwood based his studies on the spread of infectious diseases in herds of mice, he retained that "the many questions regarding epidemics can only be answered by finding out actually what happens in an infected herd, not by deducing what might happen from knowledge of what occurs in individual hosts." He made some experiments using herds of 100 000-200 000 mice and he found, for example, that it was possible to mantain for months or years herds infected with bacterial parasites such as *Salmonella typhi-murrium* and *Pasteurella muriseptica* without any cross-infection; he also made some additional experiments to observe the effect of several meth-

ods of interference on the spread of the infectious diseases that he considered. From statistical examination of results, Greenwood concluded various main aspects of the disease spread: the disease never died out in heards of mice living in close and continuous contact and considering that the herds were subject to continuous and intermittent immigration of susceptibles; the average resistance of surviving mice increased with survival in the herd but never became absolute; the selection, by death of the more susceptible or by natural immunization, had an important role in the increased resistance displayed by surviving mice; and so on. Furthermore, with his studies Greenwood tried to find some mechanisms to eradicate the disease and to immunize the herds changing the conditions of contacts and the characteristics of the spread of the infection.

Another relevant epidemic models developed in the 1920s is the *Reed-Frost model*, a mathematical model of epidemics by Lowell Reed and Wade Hampton Frost, of Johns Hopkins University [33]. This is an example of simplified, iterative model of how an epidemic will behave over time; it is based on the following assumptions: the infection is spread directly from infected individuals to others only through a certain type of contact (adequate contact); any non-immune individual in the group, after such a contact with an infective individual will develop the infection within a given period and will be infectious to others only within the following time period, while in subsequent time periods, he/she is permanently immune; each individual has a fixed probability of coming into adequate contact with any other specified individual in the group within one time period, and this probability is the same for every member of the group; the population is closed; finally, all these conditions remain constant during the epidemic. Knowing the size of the population, the number of individuals already immune, the number of infectives and the probability of inadequate contact, this model allowed to evaluate how many individuals will be infected and how many immune in the next time interval. Furthermore, repeating this model several time by changing the initial conditions it is possible to observe and evaluate how these effect the progression of the epidemic.

As well as for all innovative scientific ideas and their formulations, a crit-

ical analysis of these first epidemics models followed their exposition. In [46], M. S. Bartlett asserted that any complete quantitative theory was based on hypothetical systems or models depending on some parameters whose values could be determinated by observing; for example, in the case of infection from person to person, some of these parameters specified the nature of incubation and infectivity periods and the probability of transmission, but other variables could be considered, as the size and the structure of the susceptible population and changes in immunity to the infection. Therefore, it was of the utmost importance to accurately determine these parametric values from appropriate statistical data. Furthermore, Bartlett believed that a relevant problem was that such information by itself did not automatically lead to an understanding of the behaviour of population as a whole, and the justification of theoretical discussion was that the mathematical models of typical epidemiological situations suggested complex conseguences even on the simplest assumptions. According to his idea, the introduction of simple division of the human population in susceptible, infected, and recovered was not the only aspect on which to focus, indeed Bartlett emphasized the importance of understanding the characteristics of the invading pathogen. For example, in his critical analysis of the work developed by Ross on the formulation of the epidemiology of malaria he stated that it was necessary to include in that formulation also the characteristics of the population of mosquitoes that transmit the infection. For this reason, the classical deterministic models were not sufficient to describe the evolution of an epidemiological situation, so Bartlett suggested the use of statistical or stochastic models: in the complete study of an epidemic, neglecting random or change factor could be quite misleading. Bartlett focused on the mechanisms for recurrent epidemics, that is, when the susceptible population is in one way or another replenished, and in particular on the mathematical model for the measles, but he claimed that the theoretical equations and techniques developed were applicable to epidemics models in general.

In general the 1920s, defined the *Golden Age of Theoretical Biology*, constituted a particularly productive period for the development of those theories. The work of Kermack and McKendrick inserts in the general debate

that, after a period of great brightness, stops or proceeds slowly until the 1970s, when a new period for the mathematical modeling of epidemics opens again. Indeed, mathematical modeling of infectious diseases has progressed dramatically over the past 4 decades and continues to prospers at the nexus of epidemiology, infectous diseases research, and mathematics. Mathematical models are being integrated into the public health decision-making process more than ever before, because they are recognized as a valuable tool. An example is carried by Roy Anderson and Robert May that in 1990s consolidated concepts in mathematical epidemics and provided new insights into spread of HIV infection, [34;35;36;37]. They focused considerable attention, using the differential equations of Ross model, on the role of infectious diseases in the considered dynamics of host populations including invertebrate host. They studied the role of parasites, defined to include viruses, bacteria, and protozoans, in biological control and they determined measures necessary for the eradication especially of viral diseases such as rabies, measles, and whooping cough. In most cases, their models predict the existence of a threshold that is a consequence of the assumption that the rate of disease transmission is proportional to the number of random encounters between susceptibles and infectives in a population.

Most of the models developed have involved aspects such as passive immunity, gradual evanescence of vaccine, and disease acquired immunity, stages of infection, age structure, spatial spread, vaccination, quarantine etc. Special models have been formulated for diseases such as measles, rubella, chickenpox, smallpox, malaria, whooping cough, HIV/AIDS, etc. In [38], Becker et al. believed that only a part of the transmission process of a disease is observable; indeed, sometimes only the eventual number of cases is observed and so only certain parameters of the transmission model are estimable; e.g., neither the times of infection nor the times when the infectious periods start are observed. As a consequence, simplifying assumptions are needed. Thus, Markov chain Monte Carlo (MCMC) methods are used for the analysis of infectious disease data [86]. Two important data set were considered, containing temporal and non-temporal informations respectively, from outbreaks of measles and influenza. Their purpose was to provide some examples of the

use of the MCMC methods and to illustrate how various realistic modelling assumptions could be readily incorporated. Some preliminary work on the application of MCMC methods for simple epidemic models can be found in O'Neil and Roberts [39] and in Gibson and Renshaw [40]. Also in [39;40] the analysis of infectious disease data was usually complicated by the fact that real life epidemics were only partially observed and, in particular, data concerning the process of infection were often unavailable.

A model for a specific infectious disease is [41], in which Castillo-Chavez, based on the idea that mathematical epidemiology has resulted from the need to understand and control the global epidemic such as AIDS, focused on the implications of variable infectivity, the immune system and social/-sexual mixing dynamics for our understanding of the epidemic process: he measured the infectivity of the diseases, evaluated the stages of the infection, and the associated transmission probabilities. He also tried to model heterogeneity in susceptibility and infectivity by introducing heterogeneity parameters multiple. Addressing the immune system, he developed models to describe the complex interaction of the immune system and HIV, considering some features such as the long latency period, or the almost complete absence of free virus particles etc. As well as Castillo-Chavez, also Hethcote and Van Ark made some study and applied mathematical models to the spread of AIDS, [42]; in [43], Hethcote studied also a mathematical model for the spread of gonorrhea, he considered several background, for example the type of population for the evaluations, and compared the results. Indeed, in the 1970s gonorrhea led the list of infectious diseases in the number of cases reported by the U.S. Public Health Service, with more cases than the combined total for several other diseases, as syphilis, measles, infection hepatitis, etc., so many other models for the evaluation of the spread of this disease developed as well as [43]. The distinctive epidemiological characteristics of gonorrhea caused the models to differ from those of other infectious diseases. In [47;48] Yorke et al. designed a model with time-independent coefficients; this model based on some relevant characteristics of gonorrhea, such as the average incubation period that is short (3 to 7 days) compared to the often quite long period of active infectiousness, or the fact that often this disease

is asymptomatic and that an infected individual seems to remain infectious until he/she receives antibiotic treatment. The asymptomatic cases of gonorrhea do not seek prompt medical treatment and so these cases are infectious for periods much more longer than the latent period. Another aspect that Yorke et al. took care of was that no significant physiological immunity is derived from having previously been infected; indeed, there were individuals who have been infected and cured over and over, so they assumed that as soon as the curing antibiotics had left the body, the individual was again susceptible. Therefore, the model considered only infectives and susceptibles, without immunes; moreover, they concluded that unlike many diseases, the duration of the infection and the contact rates were extremely variable. A more or less recent work that describe instead models for tuberculosis epidemics is [74]; this is another example of a specific formulation for an infectious disease; recently tuberculosis, although preventable and curable, causes more adult deaths than any another infectious disease. The theoretical framework of [74] designes and develops effective control strategies to determine treatment levels for eradication and quantify the effects of non-eradicating control. The theoretical formulations were extended to assess how suboptimal control programs contribute to the evolution of drug resistance and the authors developed a new evaluation criterion to suggest how control strategies can be improved.

A detailed analysis of the basic model for the general evaluations of the diseases spread, the *SIR model*, is made by M. J. Keeling and P. Rohani (2008), [44]; their book is designed as an introduction to the modeling of infectious deseases: they start with the simplest of mathematical models and show how the consideration of appropriate elements of biological complexity leads to understand the disease dynamics and their control. The *SIR model* is based upon calculating the proportion of the population in each of the three classes, susceptible, infected, and recovered, and determining the rates of transition between these classes. Several variations of the *SIR model* are explained; although the basic model provides a simple and generic framework for understanding and predicting epidemiological dynamics, a number of modifications are possible, which increase the model realism but also the

number of parameters that have to be estimated to better analyze the infectious diseases. Moreover, in [44] both deterministic and stochastic models are described; the persistence of infections, particularly childhood infections, whithin a population inspired also the study of stochastic models, in which the number of individuals in all the classes was always an integer and events happened at random but with a given underlying probability that was based on the associated deterministic model.

The purpose of some of the models developed during the years is to use them as either a predictive tool or as a means to understand fundamental epidemiological processes. Indeed, one of the primary reasons for studying infectious diseases is to improve control and obtain some methods of eradication of the infection from the population: the studies of these methods and models allow to formulate several forms of control measure, all operated by reducing the average amount of transmission between infectious and individuals that are susceptibles. The control strategy to use depends on the disease, the hosts, and the scale of the epidemic. The first studies of Bernoulli [29] that led to the eventual smallpox vaccine, developed by Jenner. The purpose of vaccination is to reduce the number of suscetible individuals applied to a large proportion of the population. Other mechanisms are used to control the infection; for example, the isolation of known or suspected infectious individuals, called quarantine, is one of the oldest known form of disease control and still in use, indeed it was used to combact SARS in 2003, and it is a rapid first response against invading pathogens. The main idea is that it essentially operates by preventing infectious individuals from mixing with susceptible ones, hence stopping trasmission. However, quarantine can be applied only once an infectious individual is identified, by which time the individual may have been transmitting the infection for several days. These and other measures of diseases control are not 100% effectives in most cases. Furthermore, with the recent development of several means of transportation that become faster and faster allowing milion of people of all nationality every day to travel around the world, the spread of infectious diseases is easier. Therefore, for all these reasons, it is necessary to formulate accurate models and advanced mechanisms of control of the diseases spread.

2.2.2 Networks and SIR models

Networks and the epidemiology of directly transmitted infectious diseases are fundamentally linked. The foundations of epidemiology and early epidemic models were based on population wide random-mixing, that is populations within which interactions take place between randomly selected individuals. In practice, however, each individual in a population has a finite set of contacts to whom they can pass infection: the whole structure of all such contacts forms a *mixing network*. The knowledge of the structure of that network allows models to compute the epidemic dynamics at the population scale from the individual-level behavior of infections.

The historical study of networks begin in two different fields: social sciences and graph theory [87]. Whereas the epidemiologists talk about "hosts" and "contacts," graph theory uses the terms "nodes" and "edges," while the social literature, on which social sciences are based, talks about "actors" and "relations." In each case, it is the presence of a relationship between elements in a set of these that is the issue of concern. Research in the social sciences provides quantitative and qualitative informations about social network connections, which are related to the mixing networks for infectious diseases, because it focuses on the network connections rather than on the properties of the network itself. For example, Leinhardt, in the 1977 [49], used network analysis to describe the evolution and spread of ideas and innovations in societies and observed that social dynamics can be understood through analysis of the social networks on which they are based. Attention has been given to the nature of connections: properties such as symmetry and transitivity, which together allow to measure the social cohesion.[1] Also measures of the importance of individuals heve been derived, based on various considerations, such as the number of connections, or other structured properties [50]. Such ideas immediately inspired epidemiologist because the concept of the social importance of an individual is directly related to its role in disease spread. Also research in graph theory has provided quantitative tools and

[1] A relation is symmetric if a relationship between X and Y implies the relationship between Y and X, a relation is transitive if a relationship between X and Y and a relationship between X and Z imply a relationship between Y and Z.

mechanisms to describe networks, many of which have epidemiological applications. For example, within this research the notion of adjacency matrix, or sociomatrix, is developed to describe the connections in a population; in general, this matrix summarizes all the connections within the networks [51;52], and in epidemiology it is used to represent the transfer of the infection from an individual to another in a population.

Because epidemiology is focused on the spread of a disease, the network forms a constraining background to the studies on the transmission dynamics, in contrast, the research in graph theory and social sciences considers the understanding of the network itself as the ultimate goal. For this reason, the same tools are often used for different purposes; however, the problems that arise in the analysis of an epidemiological network are the same that complicate the study in other fields. Determining a complete mixing network requires knowledge of every host in a population and its relationship with every other host, so the amount of data required represents the first relevant problem because the collection of all these data is a time consuming task. Even when an entire population can be sampled, there are other factors that complicate network evaluation; for example, the fact that the evaluation of contacts requires personal informations may not always be volunteered, especially for sexual mixing networks. Moreover, because different infectious diseases are passed through different paths, a mixing network is necessarily disease specific; thus, a network used to describe HIV transmission would be different from one used to examine influenza.

The development of the concepts of *small-world networks* [77;78] and *scale-free networks* [79] from the initial graph theory and the idea of random networks introduces new innovative tools in the evaluation of an infectious disease spread. Considering the properties of *small-world networks*, the high level of clustering means that most infections occur locally, while short path lengths mean that the disease spread through the network is rapid and infection is unlikely to be contained within small regions of the population [53]; moreover, percolation theory is often applied to *small-world networks* to evaluate threshold parameter values at which epidemic can take place [54]. Because highly connected individuals are likely to be very important

in infectious disease transmission, considering these elements into network is necessary if network must capture all complexities of disease spread, and *scale-free networks* provide the appropriate tools for achieving such extreme levels of heterogeneity. *Scale-free networks* can be constructed by adding dinamically new elements to a network, one at a time the connection mechanism follows the rationale of preferential attachment [55;56], that is each new individual added to the population connects preferentially to individuals that already have a large number of contacts. This refers to the power-law distribution, initially observed for World-Wide Web connections, and then power-grid networks, graphs of actor collaborations and networks of human sexual contacts. For epidemics, the power-law distribution play a pivotal role in the spread and maintenance of infection. Indeed, having many contacts has two relevant effects: the individual that has several contacts is at a greater risk of infection and can transmit the infection to many others once infected. Furthermore, in the preferential attachment model the existence of individuals of arbitrarIly large connections means that there is no level of random vaccination that is sufficient to prevent an epidemic [56]. It becomes possible to control infectious diseases through vaccination when there is some upper limit in the number of contacts that an individual can have [57]. In [56], Pastor-Satorras et al. highlighted the dominant role of such individuals, because, for example, the vaccination of only a few of these can be sufficient to prevent an epidemic reinforcing the standard public-health guidelines.

Until now, the influence of several properties associated to different type of networks have been investigated, but nowadays, with the development of increasingly accurate epidemic models, it is possible to find in the literature some applications of those models to solve relevant network problems for computer science and telecommunications. An example is [58], in which an immunized SIR model is used to design a routing protocol for sparse MANETs. The topology of a MANET is constantly changing because of the movements of the mobile devices. Thus, the basic challenge is to equip each device in the network with an effcient routing protocol which allows them to maintain the information necessary to appropriately route or forward data in the network. In a MANET two nodes in coverage range of each other

are directly connected and the data can be directly delivered between them without the help of any intermediate nodes; conversely, if two nodes are not in coverage of each other, some intermediate nodes are necessary to deliver data. To solve this problem, routing protocol based on epidemic theory were shown to improve the performance, achieving better transmission rates and lower delays. Epidemic routing based on SIR model depends on forwarders of messages coming into contact with other nodes through the movements of that nodes. Therefore, the purpose in [58] is to efficently deliver the data packets in such scenario without any considerable delay and to reduce the amount of resources used in delivering the data.

Another important problem that particularly affects computer networks is the spread of viruses; for example, today the propagation of active worms is one of the most important issues. In general, a computer worm is a self-propating malicious code, and an active worm is a computer worm which find its victims in the network, especially in peer-to-peer networks, by using the vulnerability, found through scanning procedure, of that victims. Thus, the study of the propagation of that virus is an important challenge for network security. Reference [59] discussed this scenario by using a continuous time Markov chain, the SIR model has been developed as the basic model; indeed in [59] a mechanism based on the epidemiological model considering hosts joining and leaving the network is used to study the propagation of topology-aware active worms that scan the network by using the information given by the network topology, and the authors found that the hosts dynamics have a relevant impact on the size of epidemic and influence the propagation performance. Also in [60], the impact of joining and leaving hosts on the spread of topology-aware active worms is studied in peer-to-peer networks, based on SI epidemiological model. In [65], instead, the threat of virus spread in wireless sensor networks is studied. Also, the mechanism introduced by this paper is based on the SIR model, modified into the Susceptible-Infective-Recovered with Maintenance to characterize the dynamics of the propagation of the virus from an initial single node to the whole network. The name of that modified epidemiological model is due to the introduction of a maintenance in the sleep mode of a wireless sensor network and it can improve the anti-

virus capability of the network and allow the adjustment of the network to various type of virus without additional compuational or signaling overhead.

A further application of SIR model in peer-to-peer systems is described in [61]. Here the analogy of peer-to-peer networks to biological epidemic model is highlighted for file sharing: the susceptibles are the idle peers that generate file requests with a certain rate; once a peer starts to download the file, it becomes a downloading peer, that is infectious, and once the downloading is finished the peer joins the sharing peers group, that is the group of recovered.

From the above, the network version of the SIR model is based on local rules of transmission which take care of the network topology. Moreover, understanding the propagation of information on complex networks is a key issue from a theoretical and applied point of view. Indeed, in [62] it is explained how traces of peer-to-peer file sharing based on SIR model can be used to evaluate the propagation of large-scale real-world data which nowadays remains an important challenge due to the scarcity of open and extensive data.

The epidemiological SIR model and its variations have also been used to study the problem of detecting the information source in any type of network in which the propagation of information follows the popular epidemic model. This is the reverse of the diffusion problem: given a characterization of the diffusion process in a certain time t, can we tell which is the source node of the propagation? To this question an answer can be found in [63;64]. These studies carry also to answer to other more general questions, such as which computer is the first one infected by a computer virus? or who is the source of a rumor in online social networks? or, in epidemiology, where is the source of an epidemic? In [63], given a description of the network, from which all infected nodes are known but it is impossible to distinguish between susceptible nodes and recovered nodes, is developed a sample path based approach, where the estimator of the information source is chosen to be the root associated with the sample path that most likely leads to the observed characterization, to solve the problem of finding the information source based on the description and the topology of the network considered. The authors assumed that recovered nodes cannot be infected again and

that initially all nodes are susceptibles except one infected node, that is the information source that at the beginning infects its neighbors, and thus the information starts its propagation in the network. The result of this work, through the evaluation of the performance of the reverse infection algorithm, is that with high probability the distance between the actual source and the estimator is constant, independent of the number of infected nodes and the time used to observe the network. In [64] a similar study focus on the problem of detecting multiple information sources under the SIR model.

2.3 Epidemic models and Bibliometrics

The first who talked about the relationship between epidemics models and diffusion of ideas was William Goffman, who served as a researcher, Professor, Dean and Emeritus at Case Western Reserve University. In 1964, he published the first seminal paper in *Nature* [70] in which he stated that the dissemination of scientific ideas could usefully be described as a process similar to the transmission of disease. Indeed, he suggested that existing mathematical models that describe epidemic processes could be valuable tools for information scientists as well as for medical researchers. Goffman identified the main elements of epidemics. The first one is the infectious material itself, and how it is communicated: in medical epidemics it is a virus, bacterium, parasite, fungus etc.; instead, in intellectual epidemics, *ideas* are the infectious material. The second element is the population through which they spread: in medical field the members of the population belong to one of three categories (susceptibles, infectives, removals); in intellectual epidemics authors or researchers are infectives who have ideas to communicate, suceptibles are those who come in contact with the infectious material, and removeds are those who resist ideas or are no longer active researchers because of retirement or death, or also, those who abandoned this branch of study.

Later, Goffman has applied the epidemic model to the literature of mast cells to see how well it accounts for the nature of scientific growth and the spread of information [71]. He defined the basic population as the total num-

ber of authors listed in a bibliography compiled by Selye, that included all the contributions to this topic, from the discovery of mast cells in 1897 until 1963. The bibliography listed 2195 authors and 2282 publications. Goffman considered the diffusion of this literature as an epidemic process involving the direct transmission of ideas between authors, which were classified as infectives or removeds. He assumed also that authors became infectives in the first year of publication of their articles and, then, became removeds one year after the date of publication of their last paper in the Selye bibliography. Then, Goffman plotted the rates of change over time of the number of publication and the number of authors and found some interesting results. First of all, the curves indicated that changes in the number of publications mirrored those for authors, thus, the epidemic explosion of mast cell research is simultaneous to the population explosion of authors. He also studied the stability of that spread and determined which of the several separate lines of investigation in mast cells was the most virulent in terms of size and intensity.

With the development of the technological field and the institution of online databases, two recent works [72;73] have been found on the application of epidemiological models to ideas spread. Indeed, the kind of analysis such as those conducted by Goffman were more difficult and time consuming at his time. In [72] the authors applyed several epidemiological models to empirical data on the spread of Feynman diagrams through the theoretical physics communities of the USA, the URSS, and Japan after World War 2. They collected the number of authors adopting Feynman diagrams and identified the adopters of the idea, or members of the infected class, based on published discussion or uses of Feynman diagrams in the main physics research journals of each country considered. Then, they estimated the effectiveness adopting the idea in the three communities and found values for parameters reflecting both intentional social organization and lifespans of the idea. The final result was that the spread of Feynman diagrams appeared analogous to a very slowly spreading disease, with characteristic progression times of the order of years instead of days or weeks.

Unlike the work made in [72], which was based on differential equations, in [73] an individual-based weighted network model is used to describe the

spread of research on kinesin, a protein belonging to a class of motor proteins found in eukaryotic cells. The authors did not consider the simple count of number of published papers or publishing authors, but they inquired into how a research topic spreads over an existing network of disciplines. Thus, their purpose was to capture the diffusion of topics over network of connections between several disciplines, as assigned by the ISI Web of Science's classification in terms of Subject Categories (SCs). The underlying network of citations among SCs represents the knowledge flows over the map of science and the weight of a link is chosen in order to be a good indicator of the likelihood of a SC becoming research-active in a certain area given that some other related SCs are already research-active in this specific area. To analyze the spread of kinesin-related research over this network, they used the approximations given by models used in the context of the transmission of infectious diseases.

Chapter 3

Model

3.1 Epidemiological model

The epidemiological *SIR model* is based on the initial concept by Kermack and McKendrick [32] and it is used to analyze the spread of an infectious disease. According to this model, the individuals within a population are categorized as

- **S**usceptible (S) if previous unexposed to the pathogen and never infected;
- **I**nfected (I) if currently colonized by the pathogen and infectious;
- **R**ecovered (R) if the hosts have successfully cleared the infectious disease and are no longer infectious.

Furthermore, the model is based on three critical assumptions:

- the population considered is *closed*, that is demography (e.g. births, deaths and migration) is ignored;
- the *small-world property*, according to which everyone in the population can infect and/or can be infected by anyone else; indeed, the *homogeneous mixing* in the population is considered, which means that every individual interacts with everyone else with the same probability,

and, thus, possible heterogeneities according to age, space or behavioral aspects are discarded;

- the *memoryless property*, which implies that we can make predictions for the future of the system considered based solely on its present state just as well as we could know the system's full history.

The challenge now is to describe the way in which individuals move from one class to another. Only two transitions are to be considered

- the transition from S to I involves disease transmission, which is determined by three factors: the prevalence of infecteds, the population contact structure and the probability of transmission given contact; the contact between susceptible and infected individuals is necessary for a directly transmitted pathogen and we consider that anyone can contact anyone else in the population. Moreover, the likelihood that a contact between an infected and a susceptible results in transmission must be considered;

- the transition from I to R, is simpler and involves the transition of infecteds in the recovery class once they have fought off the infection; the time period that an individual spends in the I class is called "infectious period". The recovery rate γ, that is the inverse of the infectious period, is acquired as a constant and this leads to exponentially distributed infectious period.

Figure 3.1 shows the flow diagram that provides a useful graphical method of illustrating the assumptions just stated according to the *SIR model*. The diagram uses black arrows to represent the transition between the S and I classes and the I and R classes, the gray arrow instead is used to show that the level of the infectious disease influences the rate at which susceptible individuals move into the infected class.

An important parameter of this model is the *force of infection* λ, which is defined as the per capita rate at which susceptibles contract the infection. Thus, if X represents the number of susceptible individuals in class S, then

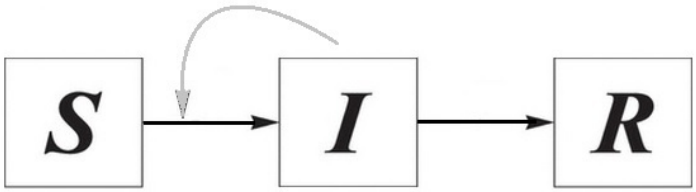

Figure 3.1: Flow diagram of SIR model

the rate at which new infecteds are produced is λX, where λ is intuitively proportional to the number of infectious individuals. According to the *frequency dependent transmission*, which refers to situations in which the number of contacts is independent of the population size, $\lambda = \frac{\beta Y}{N}$, where Y is the number of infectious individuals, N the total size of the population considered, and β is the product of the contact rate and the transmission probability, that is the *transmission rate*. Thus, it is convenient to define $S = \frac{X}{N}$ as the proportion of individuals in the population that are susceptible and $I = \frac{Y}{N}$ as the proportion of individuals in the population that are infectious. Alternatively, there is the *density dependent transmission* formulation, according to which as the density of individuals in a population increases, also the contact rate increases, and in which there is no normalization of N.

Now, we discuss the derivation of the transmission term βSI, which represents the rate at which new infectious individuals, as a proportion of the whole population, are infected, from the frequency dependent assumption. Considering an individual in the S class with an average k contacts per time unit, of these a fraction $I = \frac{Y}{N}$ are contacts with infected individuals. So, during a small time interval $[t, \quad t + \Delta t]$, the number of contacts with infected individuals is $(k\frac{Y}{N} \cdot \Delta t)$. If c is the probability of successful infection transmission following a contact, then $1 - c$ is the probability that the transmission does not occur. Thus, considering the independence of contacts, the

probability that a susceptible escapes infectious disease following $(k\frac{Y}{N} \cdot \Delta t)$ contacts, denoted by $1 - \Delta q$, is

$$1 - \Delta q = (1 - c)^{k\frac{Y}{N} \cdot \Delta t}. \tag{3.1}$$

From this, it follows that the probability that the susceptible is infected following any of these contacts is Δq.

Defining $\beta = -k \log(1-c)$ and substituting into the expression for $1-\Delta q$, it is possible to rewrite the probability of transmission in a time interval Δt as

$$\Delta q = 1 - e^{-\beta \frac{Y \Delta t}{N}}. \tag{3.2}$$

The next step is to translate Δq into the rate at which a transmission occurs, recalling that $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$, first it is necessary to expand the exponential term, then divide both sides by Δt, and thus take the limit of $\frac{\Delta q}{\Delta t}$ as $\Delta t \to 0$. The result is

$$\lambda = \frac{dq}{dt} = \lim_{\Delta t \to 0} \frac{\Delta q}{\Delta t} = \beta \frac{Y}{N} \tag{3.3}$$

which represents the transmission rate per susceptible individual. Parameter λ defined above measures the per capita probability of acquiring the infectious disease. Thus, the overall rate of transmission to the whole susceptible population is

$$\frac{dX}{dt} = -\lambda X = -\beta \frac{XY}{N} \tag{3.4}$$

and if the variables are rescaled by substituting $S = \frac{X}{N}$ and $I = \frac{Y}{N}$ in order to deal with fractions, (3.4) becomes

$$\frac{dS}{dt} = -\beta SI. \tag{3.5}$$

After outlining the basic parameters of the model used, the deterministic model equations are now introduced. In the background, a large naive population without demography is considered, into which a low level of in-

fectious agents is introduced.[1] Considering the epidemiological probabilities as a constant, the *SIR* equations are

$$\frac{dS}{dt} = -\beta SI, \tag{3.6}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{3.7}$$

$$\frac{dR}{dt} = \gamma I. \tag{3.8}$$

The purpose of these equations is to describe the development over time of the transitions between the three classes. The third differential equation (3.8) can be neglected because $S + I + R = 1$, hence knowing S and I it is possible to calculate R. Moreover, as well as all differential equations, also these ones have the initial conditions $S(0)$, $I(0)$ and $R(0)$, with $S(0) \simeq 1$, $0 < I(0) \ll 1$ and $R(0) = 0$. An example of the epidemic development generated from these equations is presented in Figure 3.2.

Despite the semplicity of this model, its differential equations cannot be solved explicitly, which means that it is not possible to obtain an exact analytical expression that represent the dynamics of S and I over the time, thus, this model has to be solved numerically. Furthermore, the analysis explained so far can also be formulated in statistical terms, according to which S, I, and R are random variables [44].

Another relevant parameter is R_0, the *basic reproductive ratio*, which specifies the average number of secondary cases arising from an average primary case in an entirely susceptible population. Its definition comes from the *threshold phenomenon*: given a population of $S(0)$ initial susceptibles, after introducing $I(0)$ infectives into the population what factors will determine whether an epidemic will occur or if the infectious disease will not spread. Considering

$$\frac{dI}{dt} = I(\beta S - \gamma), \tag{3.9}$$

[1] Because demography is not considered here, the resulting epidemic expands sufficiently quickly. Demography may change this evolution, as we will see in next section and in the final results in next chapter.

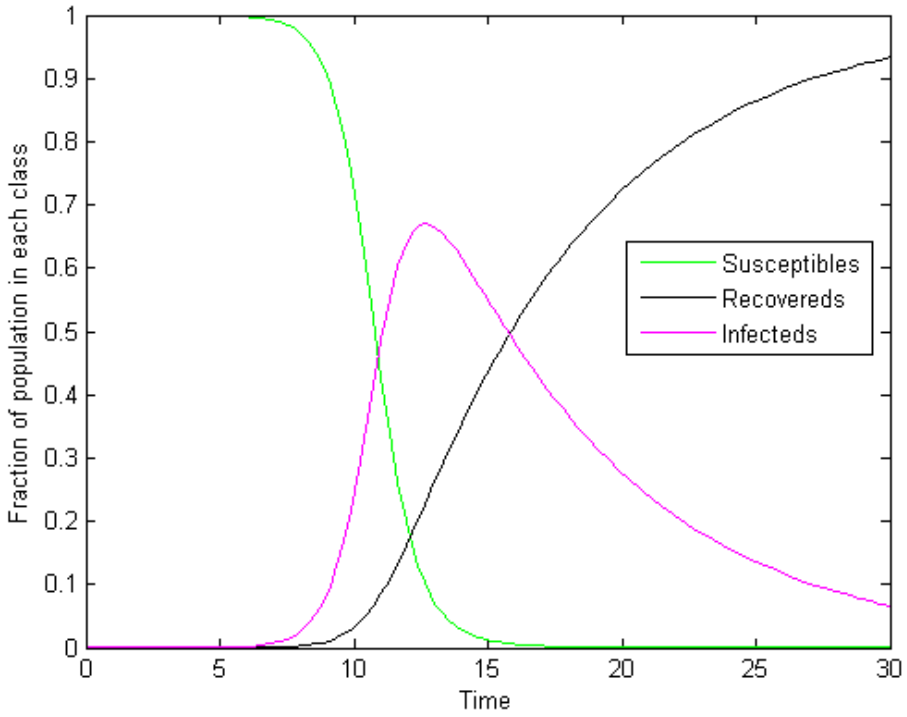

Figure 3.2: The time evolution of model variables, with an initially entirely susceptible population and a single infectious individual. It has been assumed $\beta = 1.428$ per day and $\gamma = 0.1834$ per day, giving $R_0 = 10$.

if $\frac{dI}{dt} < 0$, that is the initial fraction of susceptibles $S(0)$ is less than $\frac{\gamma}{\beta}$, the infectious disease dies out and it does not spread in the population. Thus, for an infection to invade, the initial proportion of susceptibles in the population must exceed this critical threshold. Because the inverse of $\frac{\gamma}{\beta}$ is R_0, the threshold phenomenon can be formulated in a different manner: assuming everyone in the population initially susceptible, an infectious disease can spread only if $R_0 > 1$ [44]. Intuitively, any infection that cannot successfully transmit on average to more than one new individual is not going to spread. Mathematically, R_0 can be evaluated as the rate β at which new infected individuals are produced by an infectious, when the whole population is susceptible, multiplied by the average infectious period $\frac{1}{\gamma}$.

Until now, we focused on the initial stage of the infection spread, but an important aspect can be acquired considering the long-term, or asymptotic, behavior. Dividing (3.6) by (3.8) and then integrating with respect to R with $R(0) = 0$, it can be found that the epidemic development involves an *exponential trend* due to the memoryless evolution of the disease, which leads to the decrease of the number of susceptibles and the increase of the number of recovereds, with a delay due to the infectious period. Moreover, there always will be some individuals in the population that remain in the S class without contracting the infection. This means that the chain of transmission eventually breaks due to the decline of infectives and not due to the lack of susceptibles [44]. Considering this result, and the fact that $S + I + R = 1$ and the epidemic ends when $I = 0$, it can be derived that if $R_0 < 1$ no epidemic occurs, and that whenever an infectious disease has a large basic reproductive ratio ($R_0 > 5$), more than 99% of the population is likely to contract it.

3.2 Application of SIR model to ideas spread

The infectious disease in the evaluation of a research idea is the *idea* itself, and according to the basic assumptions of the *SIR model*, it is necessary to explain them in the scenario considered. In the epidemiological background a population of individuals that are susceptible to an infection is considered;

a certain number of individuals goes from being susceptible to being infected once contracted the pathogens, and finally immune. Thus, as time passes, the number of infected individuals increases until a peak is reached, and then decreases as a consequence of the increase of the number of recovered individuals. Thus, in the time evolution of an infectious disease, if in a specific time an individual is infected, in the next time value the same individual can be infected or recovered. In the propagation of a research idea in a scientific community, the time development is quite different. The susceptibles are represented by the set of all possible articles on a specific topic that researchers can write and, then, presented at a conference, instead, the number of publications counted for a specific topic chosen represents the infecteds. Thus, the papers counted for a given month in a certain year are not included in the number of papers counted for the next month in the same year.

Moreover, although recovery is a natural concept in epidemiology because the organisms naturally may become immune after an infectious disease, when discussing the spread of an idea in term of infection spread the parallelism concerning this concept is more complex. Indeed, thinking of individuals or written articles, there is no systematic cognitive process, as well as the immune system, that clears out ideas from them.

Also interdisciplinary research activities play a role. Consider that a researcher nowadays may be extremely specialized or able to span across different fields. As a consequence, researchers can choose which topics to write articles about based on the interest that periodically each topic causes in the scientific community and which topics to neglect. Thus, recovereds could be considered as the set of all the possible works that researchers could conduct on a certain scientific topic, but they did not do. Furthermore, we can assume that in the today's scientific community, there are small distances between scientists or researchers, because their findings are published as papers in online databases.

However, all the considerations stated about the *SIR model* also apply to this scenario, the same parameters can be evaluated to observe and understand the dynamics with which scientific ideas develop and spread in the scientific community. In this context, β can be seen as the per capita idea

adoption rate and is responsible for the increase of the number of publications, γ is the inverse of the lifetime of the infection of a given idea regulates the decrease of the number of publications after reaching a peak, and λ is the rate at which infectious produce new infecteds. Also in this case, the *basic reproductive ratio* R_0 has an important role in the evaluations and in the discussion of the results, and it has the same definition given above. The same equations (3.6)-(3.8) have been used; according a *frequency dependent transmission* model, the values have been normalized over the number of total articles counted for a certain topic. Indeed, it is hard to evaluate the population size N, that could be considered as the potential of susceptible individuals that can be infected with a given scientific idea. To this end, empirical evaluations have been performed. Moreover, it is reasonable to consider the equation concerning the time evolution of the I class as a simple measure of the speed of propagation of an idea, that is simply the number of new articles published which talk about that idea.

3.3 Extended models

One important drawback of the basic *SIR model* is that it considers the immediate transition from the S class to the I class, and in the study of ideas diffusion this assumption is quite unrealistic. Indeed, some ideas require long periods of study and analysis before being put forward in written articles, and the same articles require some time to be written and then published and presented at a conference. Also scientific ideas require validations and often experiments to confirm or disprove them. Incidentally, this is similar to what happens with some diseases. Indeed, in epidemiology the process of transmission often occurs due to an initial inoculation with a small number of pathogen units; a period of time follows, in which the pathogen reproduces rapidly within the host, quite unchallenged by the immune system. During this *latent period*, pathogen abundance is too low for active transmission to other susceptibles: the individual, thus, cannot be categorized as susceptible, infected or recovered, he/she belongs to the *exposed* class. For these epidemics, the *SIR model* is extended to the so called *SEIR model*, in

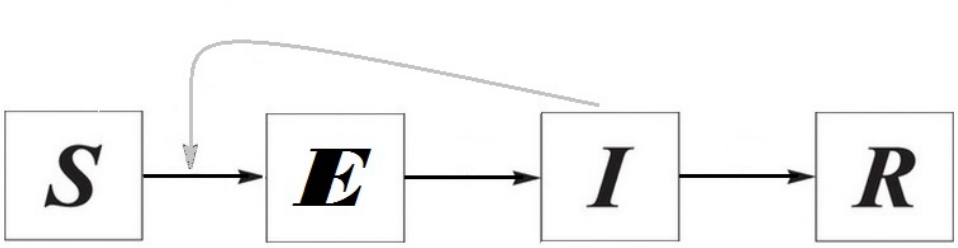

Figure 3.3: Flow diagram of SEIR model

which an additional exposed class (E class) is introduced. This model is a refinement to the *SIR model* that takes into account the *latent period*. Figure 3.3 shows the diagram of the *SEIR model* similar to Figure 3.1 for the *SIR model*.

Assuming that the average duration of the *latent period* is $\frac{1}{\sigma}$, which is memoryless as well as the *infectious period*, and, that is σ is the rate at which individuals move from the exposed class to the infectious class, the *SEIR* equations are:

$$\frac{dS}{dt} = -\beta SI, \tag{3.10}$$

$$\frac{dE}{dt} = \beta SI - \sigma I, \tag{3.11}$$

$$\frac{dI}{dt} = \sigma E - \gamma I, \tag{3.12}$$

$$\frac{dR}{dt} = \gamma I. \tag{3.13}$$

As well as the *SIR model*, also for this variation it is typically assumed $S + E + I + R = 1$ and thus the last differential equation is redundant, and the initial conditions are $S(0) > 0$, $I(0) \geq 0$, $E(0) \geq 0$ and $R(0) = 0$ ($E(0)+I(0)$ must be greater than zero for the infection to spread). The result expected from this modification of the model is that considering the *latent period* essentially could slow the dynamics of the system without actually

diminishing the extension of contagion. Indeed, the dynamic properties of the *SEIR model* are qualitatively similar to those of the *SIR model*, with the difference that the *SEIR model* has a slower growth rate after pathogen invasion due to the additional step through the E class before transmitting the infecion in the population. In next chapter, the improvement carried by this model will be shown considering some topics for which the simple basic *SIR model* is not sufficient to approximate at best the experimental developments.

Previously the basic *SIR model* has been described given that the time scale of the spread is sufficiently fast so as not to be affected by population demography. If there is interest in exploring the longer-term persistence and dynamics of an infectious disease, as well as the propagation of a scientific idea, then demographic processes will be relevant [44]. The most important factor necessary to evaluate the propagation in this scenario is the influx of new susceptibles in the population, e.g. through births of individuals that have no prior contact with the disease.

In epidemiological terms, the most common way of introducing demography in the *SIR model* is to assume that there is a natural host "lifespan", $\frac{1}{\mu}$ years. Thus, μ is the rate at which individuals suffer natural mortality; it is important to underline that this parameter is independent of the disease and is not intended to represent the pathogenicity of the infectious agent. Moreover, it has been usually assumed that μ also represents the population's birth rate, ensuring that total population size does not change through time, that is $\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$, and thus allowing stability. According to all these aspects, the *generalized SIR model* is described by the following differential equations:

$$\frac{dS}{dt} = \mu - \beta SI - \mu S, \tag{3.14}$$

$$\frac{dI}{dt} = \beta SI - \gamma I - \mu I, \tag{3.15}$$

$$\frac{dR}{dt} = \gamma I - \mu R. \tag{3.16}$$

The parameters have the same definition as (3.6)-(3.8), but it is useful

to discuss the new expression for R_0. Looking at equation (3.15), β is the transmission rate per infective and the negative terms suggests that each infectious individual spends an average $\frac{1}{\gamma+\mu}$ time units in the I class. Thus, the infectious period is reduced due to some individuals dying while infectious. Then, if the entire population is assumed susceptible, the average number of new infections per infectious individual is

$$R_0 = \frac{\beta}{\gamma + \mu}. \tag{3.17}$$

In general, this value is similar to, but always smaller than, R_0 for the case in which demographic processes are not considered, because the natural mortality rate reduces the average time an individual is contagious [44].

The development and diffusion of research ideas in the scientific community depend on the interest that this idea caused; thus, it is reasonable to focus on the number of articles that periodically are written about that topic. As a consequence, we consider μ as the parameter that represents the rate at which the number of articles per month written on this topic increases, which can be seen as a kind of "demographic birth rate." Incidentally, consider that our model already enlists a way out of the system, the transition toward state R, after which the articles are no longer tracked. For this reason, we do not need to explicitly set a death rate as well. Later, it will be shown that the simple consideration of the birth rate of the publications increase the similarity between the theoretical development and the experimental one.

Chapter 4

Results and comments

4.1 SIR model results

To evaluate and discuss the spread of research ideas in the scientific community, some scientific topics have been chosen mostly related to recent scientific trends in ICT and networking research. We collected several data about papers published in this field using the database of the *Institute of Electrical and Electronic Engineers* (*IEEE*), an international association of scientists with the purpose of promoting technological sciences [75]. The number of publications for each topic has been counted for a total of around 10 000 articles counted. In some cases, it has been considered also the analysis based on the number of publications counted using *Association for Computing Machinery* (*ACM*) [80], another database similar to *IEEE*, to verify the results found. For the sake of evaluation, we considered 8 specific subjects: `BigData`, `CloudComputing`, `SoftwareDefinedNetworking`, `InternetOfThings`, `Bluetooth`, `GameTheory`, `LTE-advanced`, and `DVB-T`.[1] We choose `BigData`, `SoftwareDefinedNetworking`, and `InternetOfThings`

[1] In particular, `DVB-T` is the acronym for *Digital Video Broadcasting-Terrestrial*, while `LTE` is the *Long Term Evolution* of UMTS and they are both telecommunications standards for cellular networks and video broadcasting, respectively; `Bluetooth` is also a standard for personal communication. `InternetOfThings`, `BigData`, and `CloudComputing` are new networking and data analysis paradigms that have recently gained popularity in the community. Finally, `GameTheory` is a subject of applied mathematics, which has found recent application in telecommunications networks and distributed systems.

because they are academic subjects and they are not related to particular technologies. On the other hand, the choice of `CloudComputing` and `LTE-advanced` is due to the fact that they are in particular based on technologies, which are more general and adaptable than those on which `Bluetooth` and `DVB-T` are based on. Finally, the analysis of the spread of `GameTheory` is interesting given its interdisciplinarity. In this regard, it must be considered that in the *IEEE* database there are only publications for the engineering field. A five years time period from the beginning of 2008 to the end 2012 has been considered; only articles that have been presented in conferences have been counted and those publications have been organized and distinguished according to the month of publication for each year. Two different mechanisms of counting have been used:

- counting the number of publications that contain the name of the topic in the title (denoted as "*document title*");
- counting the number of publications that contain the name of the topic in the keywords chosen by authors of the articles (denoted as "*author keywords*").

In this way, comparing the results obtained from the two analysis it is possible to verify the statistical consistency of these results. Indeed, as we will see, the values of the parameters slightly change because the slopes of the curves change due to the different number of publications counted, even though this number is of the same order of magnitude for each topic in the two different counting mechanisms.

The collected data present several noise effects. For example, a *seasonal noise* and a *granular noise* can be identified. The former involves the periodicity of certain conferences, more suitable for the dissemination on a given topic. It may be reasonable to assume that researchers concentrate their efforts to publish at these conferences, instead of spreading their activity evenly throughout the year. Also, for academic researchers, periodic activity of semester teaching may play a role. Granular noise is also present since the number of papers is an integer value, while the normalized model obtains a continuous trend. The discretization may become relevant especially for

the initial periods of spreading, when the number of articles is relatively low. To obtain tractable data, we therefore applied standard digital processing techniques to the traces. Specifically, we windowed (most of the times by using a simple rectangular window, but at times we applied an Hamming window) the Discrete Fourier Transform of the time series, so as to keep the samples at the lowest frequencies, which ultimately give the general trend, and filter out all noises, and then we anti-transformed back the result to obtain again a time trace. An example is given in Figure 4.1 for the trace about `CloudComputing`. The blue line represents the modulus of the Discrete Fourier Transform and the peaks that we can see at 9, 18 etc., are harmonics of the seasonal noise; instead, the red line identifies the window used.

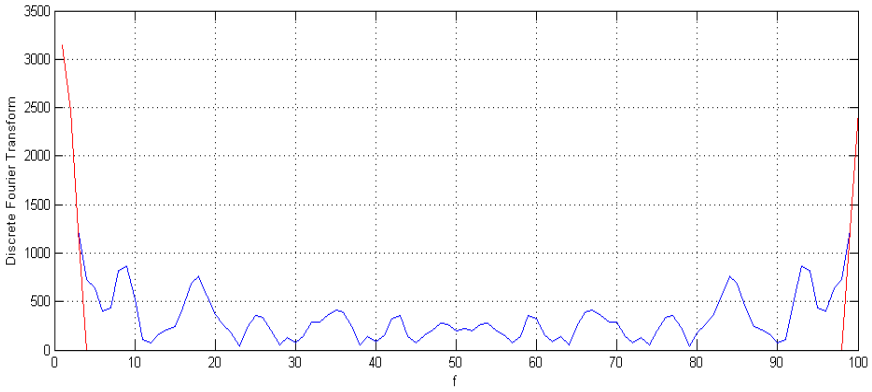

Figure 4.1: Modulus of the Discrete Fourier Transform for `CloudComputing` (*author keywords*).

We stress that this processing does not alter the total number of papers, it just spreads them more evenly, so that the peaks due to big conferences and the gaps due to the absence of conferences in a given month are smoothened out.

Table 4.1 shows the total number of articles counted for each topic; it can be observed that some topics have a significant different order of magnitude, for example, `CloudComputing` and `SoftwareDefinedNetworking`.

In general, Figures 4.2-4.5 compare the exact number of papers counted month by month for some of the topics considered, as the results found for the count based on *document title*, with the data obtained after filtering. The

former is indicated with the blue line, while the latter is represented with the magenta one.

Scientific topic	Number of papers "*document title*"	Number of papers "*author keywords*"
`LTE-advanced`	389	184
`CloudComputing`	1816	3145
`GameTheory`	577	1119
`SoftwareDefinedNetworking`	63	104
`InternetOfThings`	534	660
`Bluetooth`	474	490
`BigData`	165	123
`DVB-T`	175	131

Table 4.1: Number of papers counted.

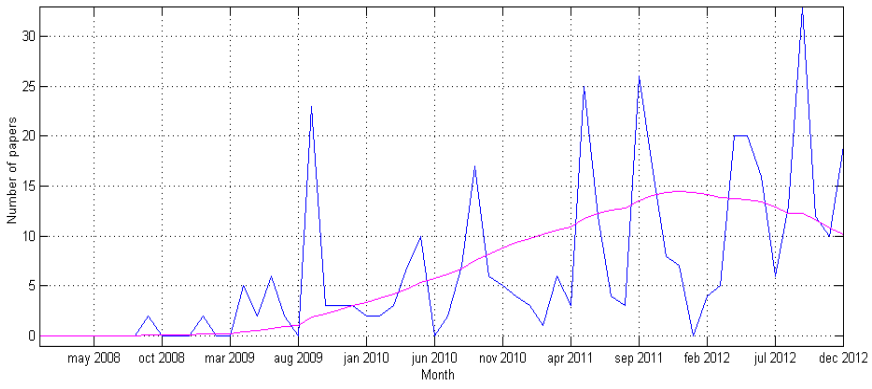

Figure 4.2: Number of publications for `LTE-advanced` (*document title*).

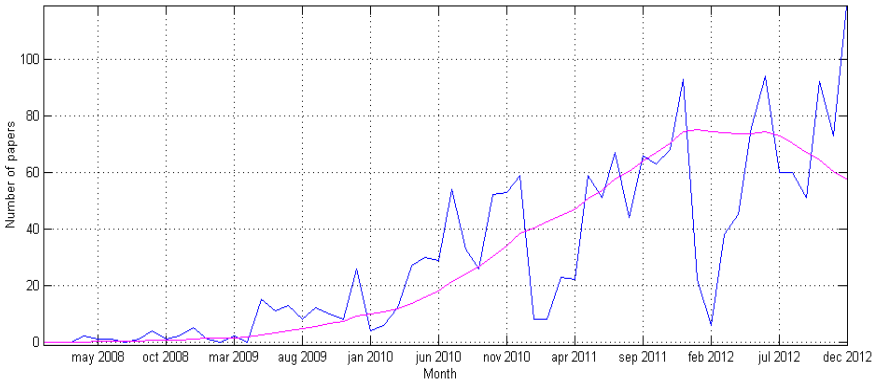

Figure 4.3: Number of articles for `CloudComputing` (*document title*).

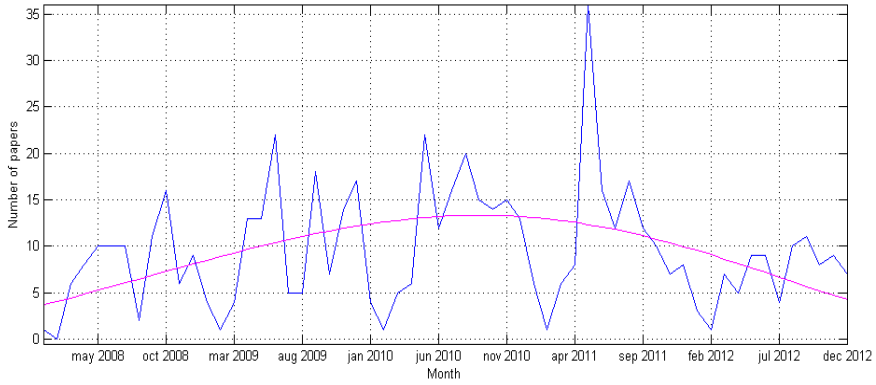

Figure 4.4: Number of articles for `GameTheory` (*document title*).

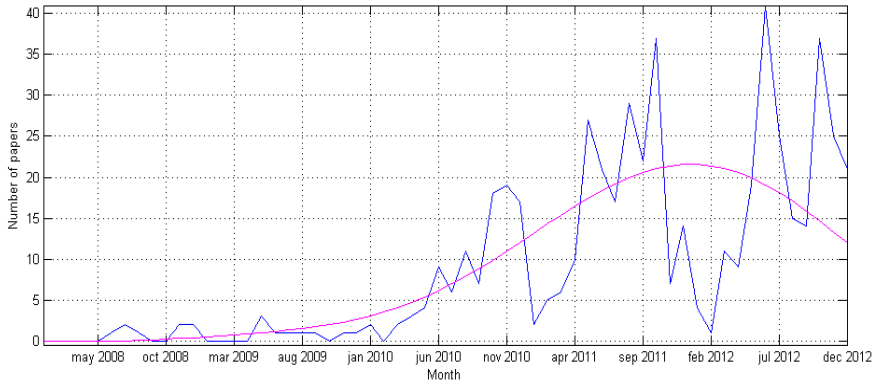

Figure 4.5: Number of articles for `InternetOfThings` (*document title*).

Moreover, scientific topics with different time dynamics and which begin their development in various periods of the given time interval have been

considered in order to capture the several aspects in the application of the epidemiological model.

Table 4.2 shows the values found for the *basic reproductive ratio* R_0, defined as the product between the *transmission rate* and the *infectious period*, for each scientific topic evaluated. Firstly, it can be observed that all the values found are greater than 1; thus, according to the parallelism with the infectious disease spread, it means that all the considered topics spread epidemically in the scientific community with an "infective" behavior. Not all the values are the same, because some topics cause more interest in the community than others. Considering Table 4.2 column R_0

R_0 *for scientific topic*	"*document title*"	"*author keywords*"
`LTE-advanced`	1.87	1.43
`CloudComputing`	1.83	1.50
`GameTheory`	1.23	1.25
`SoftwareDefinedNetworking`	1.82	2.15
`InternetOfThings`	1.42	1.43
`Bluetooth`	1.36	1.28
`BigData`	2.32	2.54
`DVB-T`	1.29	1.37

Table 4.2: Values of R_0.

"*document title*", it can be noted that the most contagious ideas are about `BigData`, `SoftwareDefinedNetworking`, `CloudComputing`, `LTE-advanced`, and `InternetOfThings` with highest values of R_0. Moreover, `GameTheory`, `DVB-T`, and `Bluetooth` have a similar (low) level of contagion; it means that, according to the results, these topics are less contagious.

In general, the comparison with R_0 "*author keywords*" shows that, even though we consider a different "population" for the counting of publications for the same scientific topic, the behavior essentially do not change significantly and the classification based on the level of contagion is more or less the same. Furthermore, this is a first way to show that some scientific topics are

popular, and very well known in and mentioned by the scientific community. For these topics, R_0 is significantly above 1. Other topics may appear as less contagious because they require a deeper and more articulate background. Thus, the simple contact with the idea itself is not sufficient to establish a solid scientific production. Also, note that for all the considered topics the values of R_0 are above 1 but not extremely high (always below 2.5). This probably reflects that the contagion of a scientific idea is a more gradual process, which cannot have the strength of a disrupting epidemic. Indeed, also validation and approval by scientific peers is required to disseminate an idea.

A special discussion must be made considering the value of R_0 of a scientific topic and the related value of the total number of articles counted for it. In particular, as we can see from Table 4.1, most topics have the same order of magnitude and the same can be said for the same topic considering "*document title*" and "*author keywords*." But, for example, in the case of `CloudComputing` compared with `BigData` and `SoftwareDefinedNetworking`, it can be observed that the total number of articles counted is quite different. This could mean that some topics (`BigData` and `SoftwareDefinedNetworking`) are more niche than the other (`CloudComputing`), that is they have an higher level of contagion, but they develop only in a narrower circle of scientists.

Now we discuss the meaning of the values found for the other main parameters and show the application of the *SIR model* through several comparison between the theoretical dynamics, obtained with the models implementation throught Matlab [44], and the experimental development. Table 4.3 and Table 4.4 summarize the results.

Table 4.3 [2] shows the numerical values of the *transmission rate*, β, the *recovery rate*, γ, and the values for the initial conditions in the count for *document title*, instead Figures 4.6-4.13 give a graphical comparison between the experimental results and the theoretical trends, represented by the dotted red line and the solid blue line, respectively. First of all, it can be noted that

[2] Due to the noise in the original data, we can consider these value to be up to the second decimal digit. The same consideration can be made for the other parameters values shown in following tables.

the theoretical results approximate well enough the experimental data, expecially for some scientific subjects such as `LTE-advanced`, `InternetOfThings`, `CloudComputing`, `BigData`, and `SoftwareDefinedNetworking`.

Scientific topic	β	γ	$I(0)$
`LTE-advanced`	0.61	0.33	$0.22 \cdot 10^{-3}$
`CloudComputing`	0.59	0.32	$0.13 \cdot 10^{-3}$
`GameTheory`	0.42	0.34	$0.55 \cdot 10^{-2}$
`SoftwareDefinedNetworking`	1.12	0.61	$< 10^{-10}$
`InternetOfThings`	0.57	0.40	$0.12 \cdot 10^{-3}$
`Bluetooth`	0.38	0.28	$0.95 \cdot 10^{-2}$
`BigData`	1.46	0.63	$< 10^{-10}$
`DVB-T`	0.31	0.24	$1.03 \cdot 10^{-2}$

Table 4.3: Values of *SIR model* parameters for the anlysis based on *document title*.

Looking at Table 4.3, it can be observed that some of the topics that previously have been found to be the most contagious, `SoftwareDefinedNetworking` and `BigData`, have the highest values of β. Moreover, considering the inverse of γ as the *infectious period*, they have also the lowest values for $\frac{1}{\gamma}$. This means that these topics are very contagious and for a limited time interval cause a relevant interest. Instead, other topics, such as `LTE-advanced`, `CloudComputing`, and `InternetOfThings`, which are among the most contagious, have a small value of the *transmission rate*, but a greater value for the *infectious period*, i.e. they have an effect for a longer time period. Furthermore, as can be seen from Figures 4.6-4.8, there is a remarkable match between the epidemiological trend and the collected measurements for `BigData`, `SoftwareDefinedNetworking` and `InternetOfThings`.

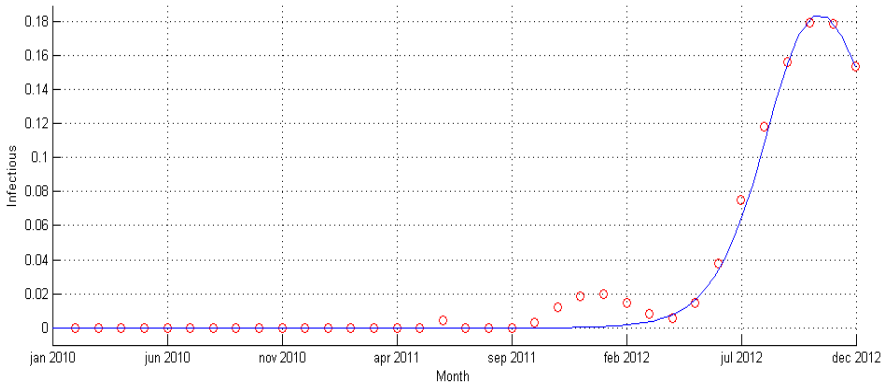

Figure 4.6: Comparison with the theoretical model for `BigData`.

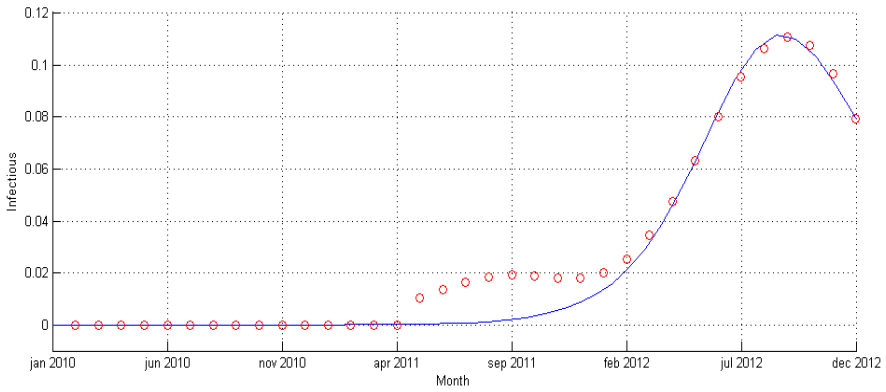

Figure 4.7: Comparison with the theoretical model for `SoftwareDefinedNetworking`.

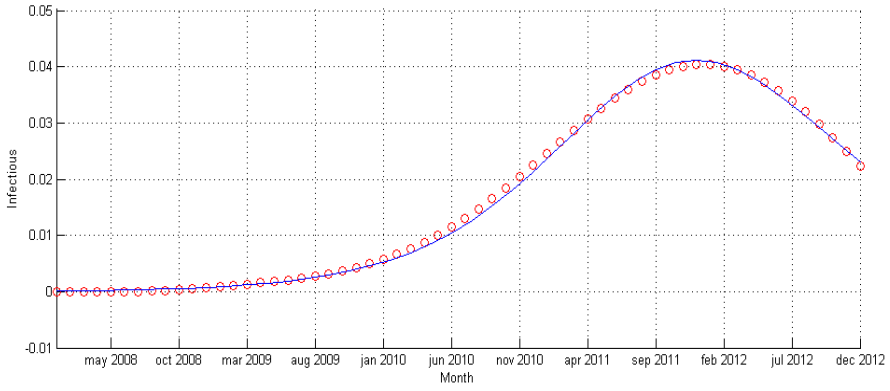

Figure 4.8: Comparison with the theoretical model for `InternetOfThings`.

It can be speculated that these subjects spread across the scientific community very much like a standard epidemics. Thus, the simple *SIR model* is sufficient to capture their dynamics. Instead, in the other cases, such as `CloudComputing` and `LTE-advanced`, we see a good match that, however, could be improved further. The reason may be that the former are more speculative topics, thus they can be considered infectious without memory; instead the others are not only academic topics, but they develop also in the techonological field, so it is necessary to consider a time interval in which the basic techonology is studied, analyzed and acquired. Indeed, we will see that the *SEIR model* gives quite an improvement.

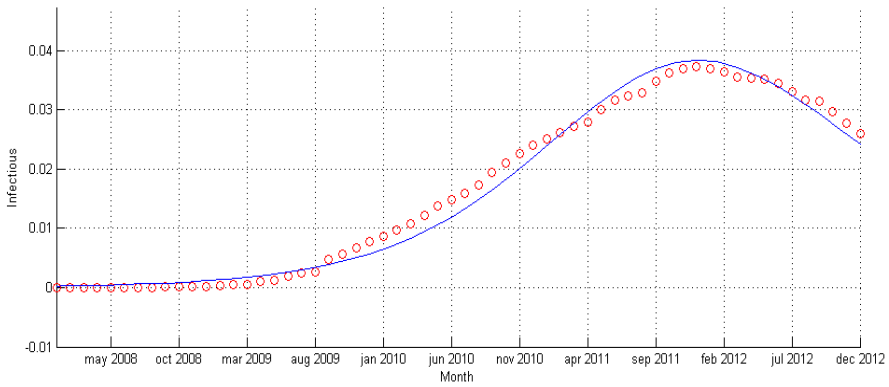

Figure 4.9: Comparison with the theoretical model for `LTE-advanced`.

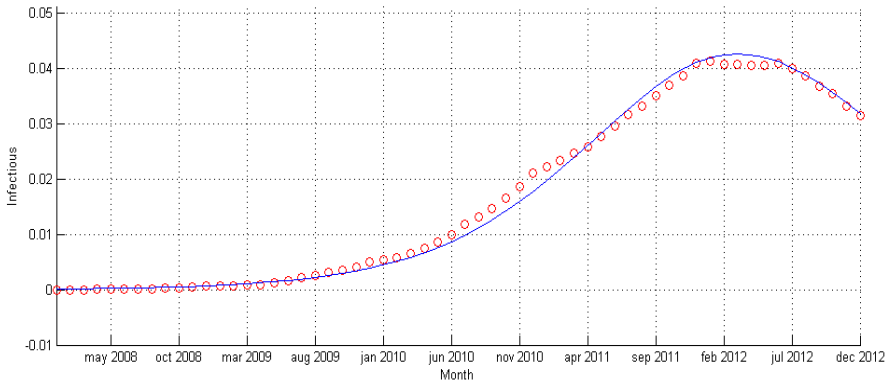

Figure 4.10: Comparison with the theoretical model for `CloudComputing`.

For the less contagious topics, `GameTheory`, `Bluetooth`, and `DVB-T`, it can be stated that in general the *transmission rate* is less than the other

topics and the *infectious period* has similar values to those of `LTE-advanced` and `CloudComputing`. This results in a lower R_0. Moreover, as shown by the figures, these topics develop before the time interval chosen and the epidemiological curves do not have a very pronounced slope. In general, these are three cases in which the simple *SIR model* alone does not work very well, in particular for `Bluetooth` and `DVB-T` for which there are maybe other non-epidemic trends superimposed. Instead, for `GameTheory`

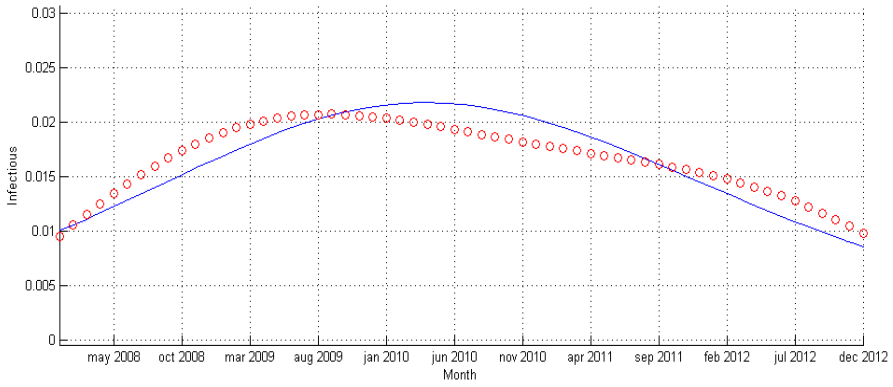

Figure 4.11: Comparison with the theoretical model for `Bluetooth`.

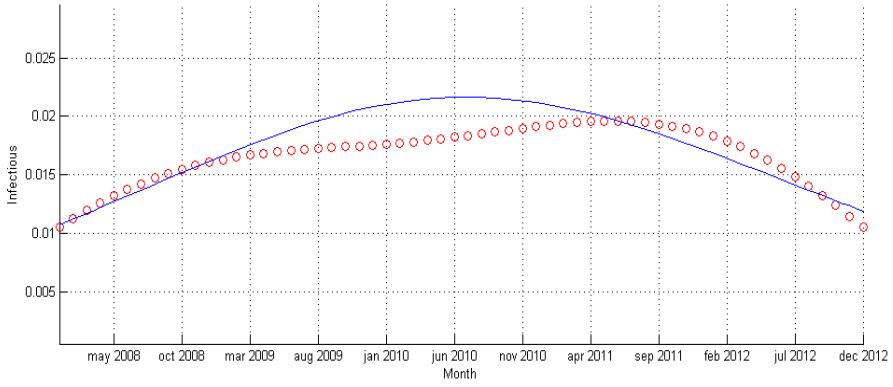

Figure 4.12: Comparison with the theoretical model for `DVB-T`.

the motivation might be found in its interdisciplinarity. It develops in several disciplines, such as mathematics and economics, thus the diffusion of that idea could be better approximate with more complex models rather than *SIR* or *SEIR* models. It may be necessary to consider a model with memory and with several vectors to approximate well the experimental dynamics.

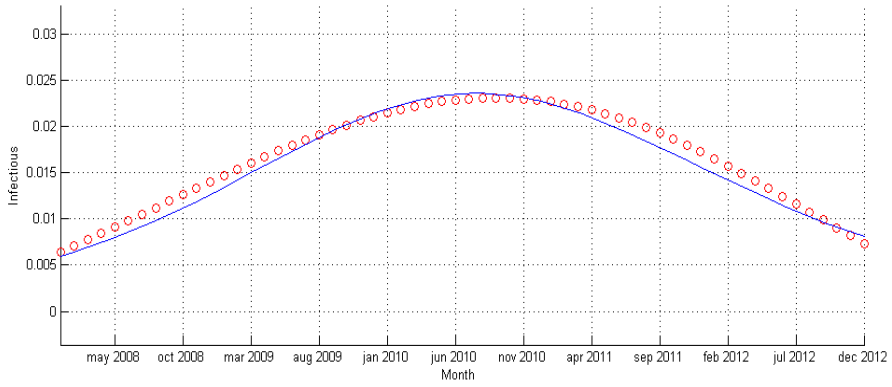

Figure 4.13: Comparison with the theoretical model for `GameTheory`.

Looking at Table 4.1, it can be observed that the number of articles counted changes considering *document title* or *author keywords*: in most cases, the total number of publications is greater for *author keywords* than for *document title*, but the dynamics of the filtered results are almost the same, as shown in Figure 4.14, where the solid magenta line represents the experimental dynamics based on *document title*, instead, the cyan one on *author keywords*, and the dotted red line identifies the filtered development based on *document title*, while, the blue one on *author keywords*. All trends shown in Figure 4.14 are related to `InternetOfThings`.

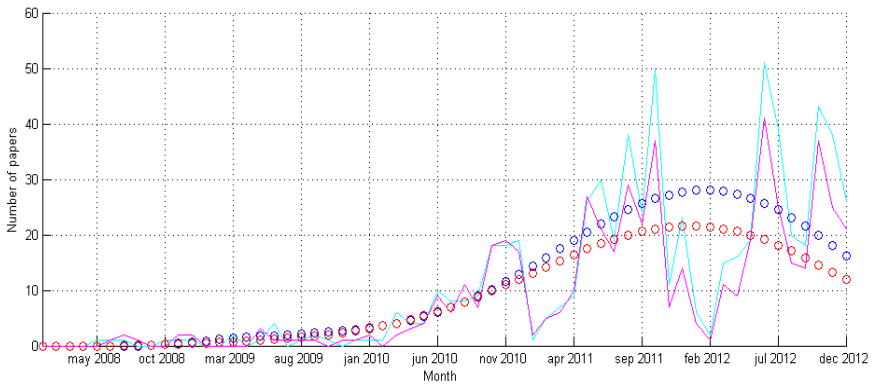

Figure 4.14: `InternetOfThings`.

Figure 4.14 shows that the filter used is similar, and also the general trend is. These remarks are also similar for other results related to the other topics considered but not shown here for the sake of brevity. Figures 4.15-4.22 give

the graphical comparison between the experimental results and the theoretical dynamics for the count based on *author keywords*, again, represented by the dotted red line and the solid blue line, respectively. Furthermore, Table 4.4 shows the numerical values of the parameters. The comparison between

Scientific topic	β	γ	$I(0)$
`LTE-advanced`	0.54	0.37	$0.19 \cdot 10^{-3}$
`CloudComputing`	0.56	0.37	$0.18 \cdot 10^{-3}$
`GameTheory`	0.36	0.29	$0.61 \cdot 10^{-2}$
`SoftwareDefinedNetworking`	1.39	0.64	$< 10^{-10}$
`InternetOfThings`	0.57	0.39	10^{-4}
`Bluetooth`	0.33	0.26	$0.73 \cdot 10^{-2}$
`BigData`	1.61	0.63	$< 10^{-10}$
`DVB-T`	0.29	0.21	$0.93 \cdot 10^{-2}$

Table 4.4: Values of *SIR model* parameters for the analysis based on *author keywords*.

Table 4.3 and Table 4.4 does not give significant differences; for some topics there is a slight increase for the value of β, for some others instead there is a decrease due to the changes in the curve slope. The same it can be noted for the values of γ. We can say that in the most extreme cases, the values are almost equal with an average variation of around 10%, but these changes are limited due to the presence of some noise and considering the change in the number of articles counted. As a consequence, we can see slight modifications in the value of R_0, but the classification based on the level of contagion for each topics is still the same, the dynamics are almost equal, and the same conclusions previously discussed can be drawn. Moreover, it is important to highlight that data given by the two kinds of count are different, even if the database used is the same and, thus, some articles are counted in both cases. The two criteria provide a list of paper partially overlapping; the overlap in certain cases is over 50% but still below 75%; in some cases the overlap percentage is even lower.

Considering the overall results found by applying the basic *SIR model*, we can state that for some scientific topics the model can be used to predict the future development and spread of the idea. This is the case of `InternetOfThings`, indeed, for this one the quantitative description given by the model matches the experimental dynamics very well. In some other cases the *SIR model* gives only a qualitative description that could be acceptable or not fully satisfactory.

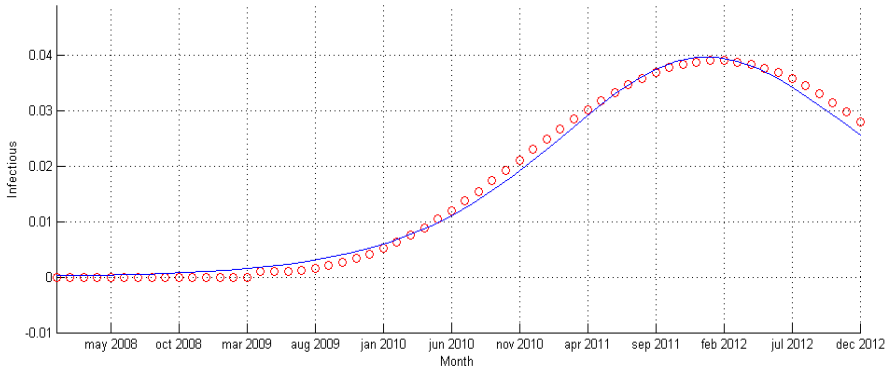

Figure 4.15: Comparison with the theoretical model for `LTE-advanced`.

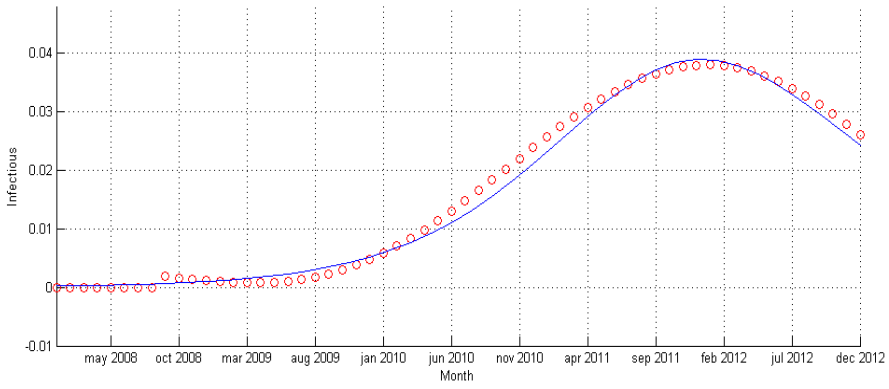

Figure 4.16: Comparison with the theoretical model for `CloudComputing`.

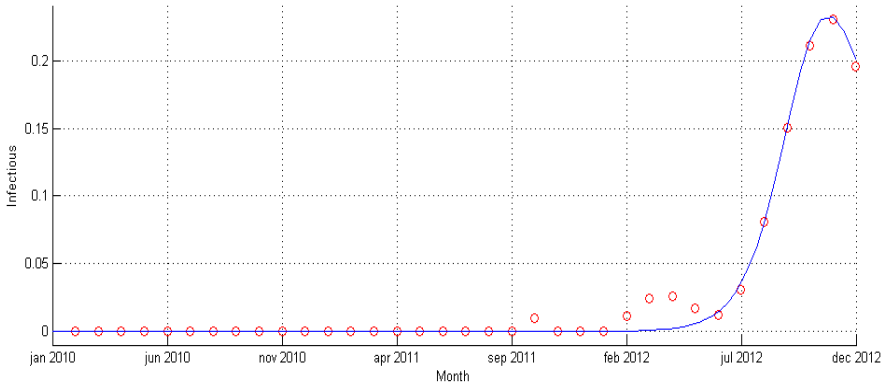

Figure 4.17: Comparison with the theoretical model for `BigData`.

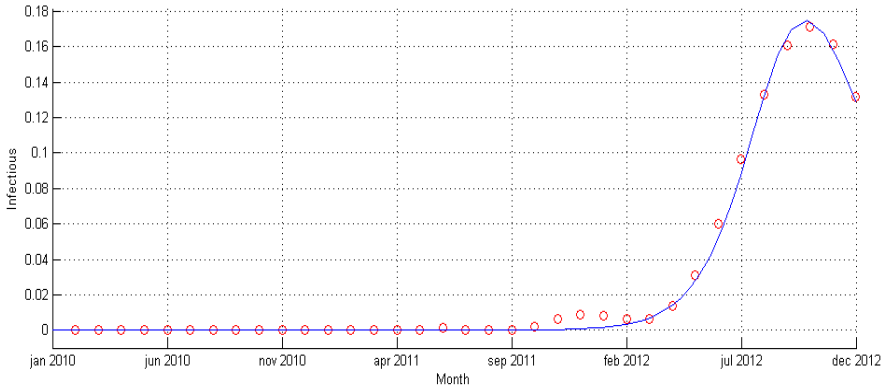

Figure 4.18: Comparison with the theoretical model for `SoftwareDefinedNetworking`.

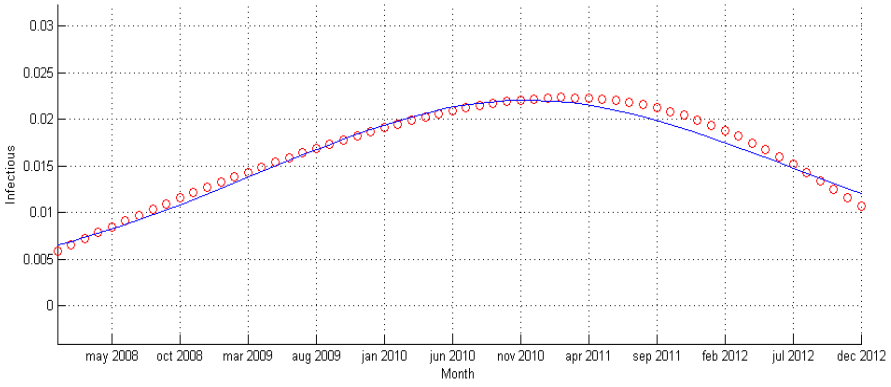

Figure 4.19: Comparison with the theoretical model for `GameTheory`.

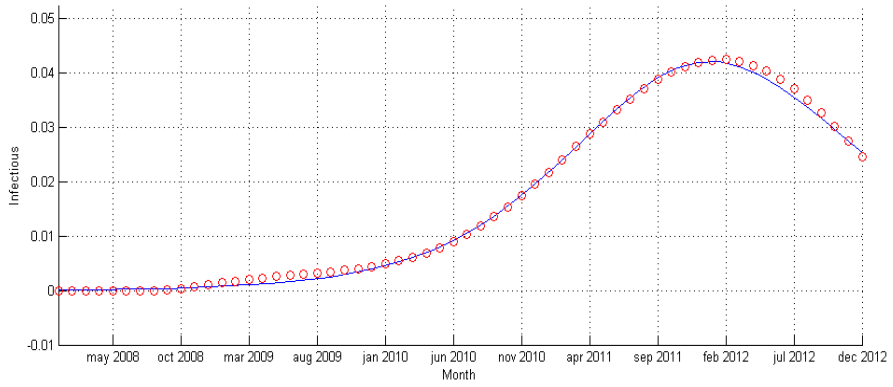

Figure 4.20: Comparison with the theoretical model for `InternetOfThings`.

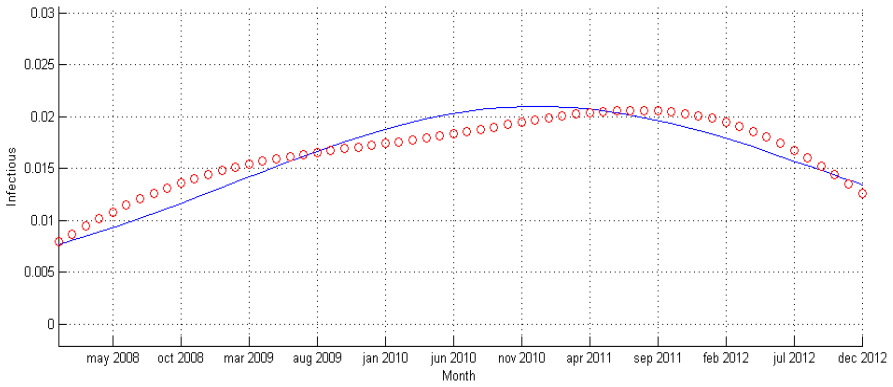

Figure 4.21: Comparison with the theoretical model for `Bluetooth`.

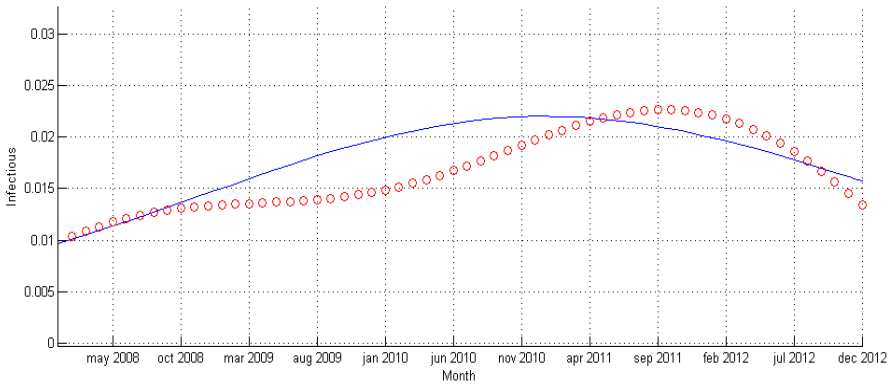

Figure 4.22: Comparison with the theoretical model for `DVB-T`.

4.2 Comparison with other databases

Comparing the different values of the number of publications counted per month, we have noted that the values found for `SoftwareDefinedNetworking` and `BigData` are significantly lower than the others. For this reason, and to verify the results found in these two cases, a similar analysis has been made based on the number of articles published by *ACM* for these scientific topics. The numbers of the overall articles counted are almost the same and also the dynamics are similar. Table 4.5 shows the values of R_0 that are similar

Scientific topic	R_0 ACM	R_0 IEEE
`SoftwareDefinedNetworking` (document title)	2.04	1.82
`SoftwareDefinedNetworking` (author keywords)	1.96	2.15
`BigData` (document title)	2.22	2.32
`BigData` (author keywords)	2.37	2.54

Table 4.5: Comparison between ACM results and IEEE results for `BigData` and `SoftwareDefinedNetworking`.

enough to verify the conclusions discussed previously. Indeed, the comparison between the results found with the two databases guarantees that the results obtained are significant. More in general, it can be conjectured that also for some other scientific topics could have similar trend across different databases. Moreover, in these cases, the databases are completely disjoint. Figures 4.23-4.26 compare both the experimental data and the theoretical development found with *ACM*.

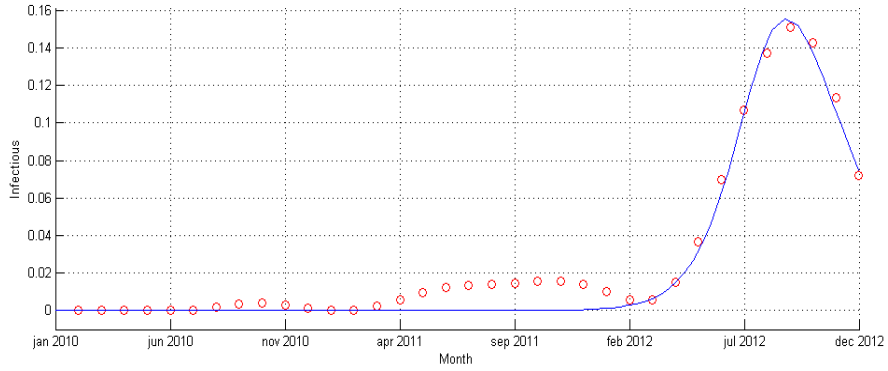

Figure 4.23: Comparison with the theoretical model for `BigData` (*document title*).

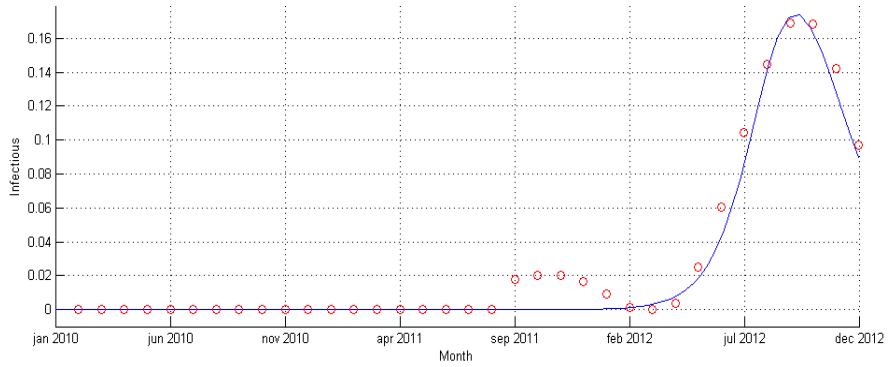

Figure 4.24: Comparison with the theoretical model for `BigData` (*author keywords*).

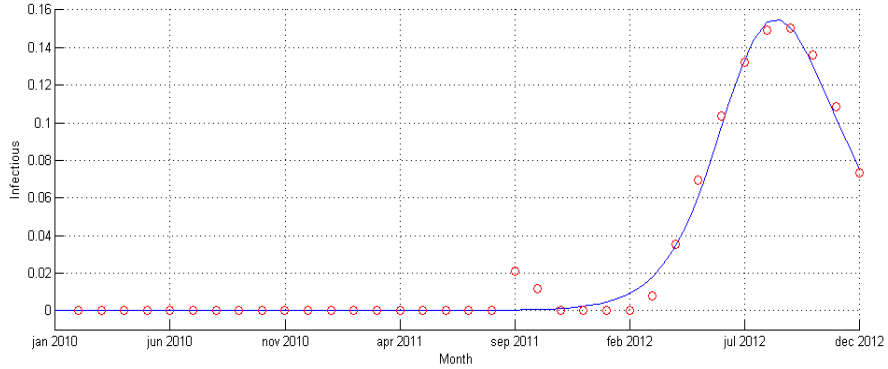

Figure 4.25: Comparison with the theoretical model for `SoftwareDefinedNetworking` (*document title*).

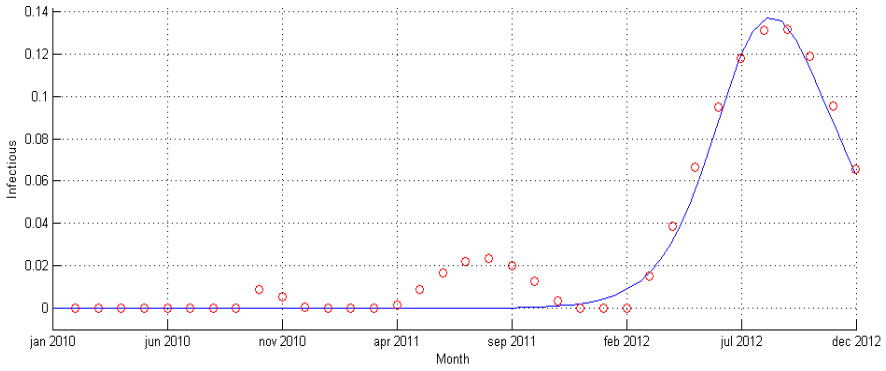

Figure 4.26: Comparison with the theoretical model for `SoftwareDefinedNetworking` (*author keywords*).

4.3 SEIR model results

We now consider a further refinement achieved by appling the *SEIR model* (see Section 3.3) instead of the *SIR model*. This implies that further *exposed* rate is added, which mimics the latent phase of an idea development. Figure 4.27-4.31, show the results for *document title*. It can be observed that, with the application of the *SEIR model*, the theoretical dynamics is much more similar to the experimental development compared to the plain *SIR model*. With this improvement, we may evaluate another parameter, apart from the *transmission rate* and the *recovery rate*, the *incubation rate* σ. Tables 4.6-4.7 show the numerical results.

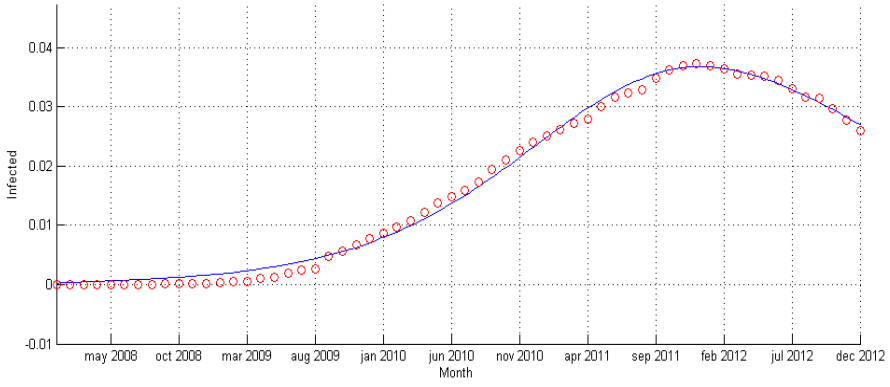

Figure 4.27: Comparison with the theoretical model for `LTE-advanced`.

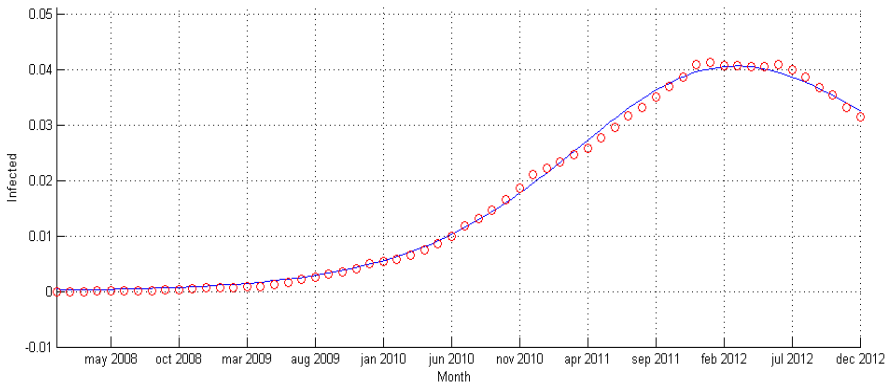

Figure 4.28: Comparison with the theoretical model for `CloudComputing`.

As can be noted, the application of the *SEIR model* might be sufficient to give a quantitative description of the experimental data in the case of `LTE-advanced` and `CloudComputing`. Instead, for the remaining three cases, `GameTheory`, `Bluetooth` and `DVB-T`, the descriptions are better than those given by *SIR model*, but there is still some margin for improvement. As a consequence, a more articulate model is necessary to describe their dynamics, in particular for `Bluetooth` and `DVB-T`.

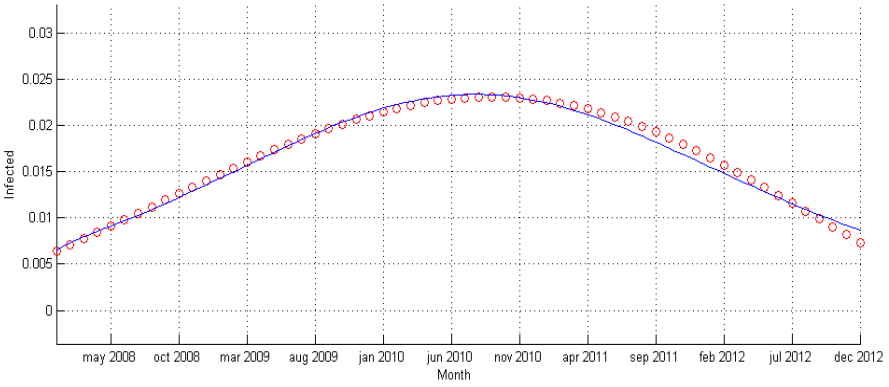

Figure 4.29: Comparison with the theoretical model for `GameTheory`.

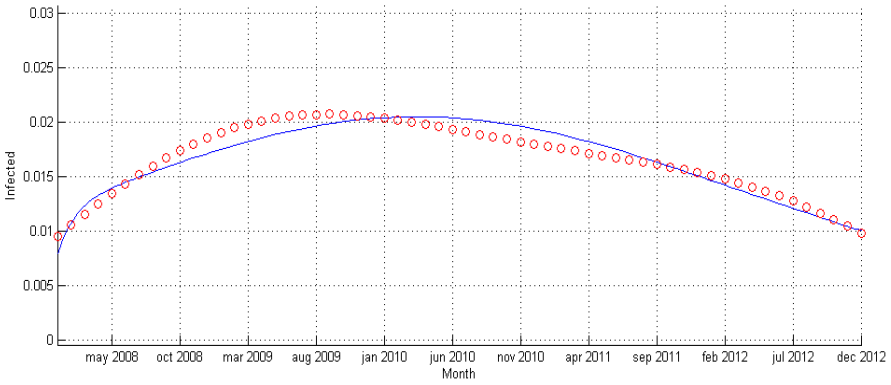

Figure 4.30: Comparison with the theoretical model for `Bluetooth`.

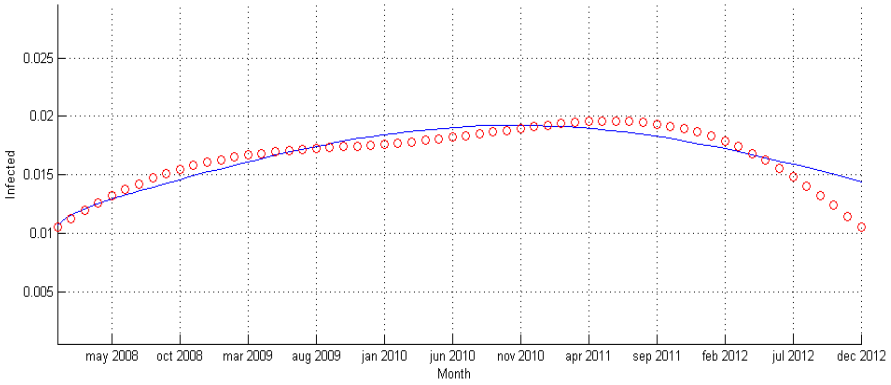

Figure 4.31: Comparison with the theoretical model for `DVB-T`.

Compared with the *SIR model* parameters, higher values for β and γ have been found for most of the topics considered, in particular the more relevant increases are for `LTE-advanced` and `CloudComputing`.

Regarding the values of σ, which represent the rate at which individuals move from the E class to R class, and thus for the *latent period* $\frac{1}{\sigma}$, it has been found that `CloudComputing` and `LTE-advanced` have the highest values for $\frac{1}{\sigma}$ (around 9/10 months). This may relate to the criterion based on which we have chosen the 8 scientific topics; indeed, we have already said that `CloudComputing` and `LTE-advanced` are based on technologies that are adaptables and related to more general concepts. For example, in the case of `LTE-advanced` the scientific community firstly had to discuss in what manner mobile wireless technologies could be extended to a new generation of cellular

Scientific topic	β	σ	γ
`LTE-advanced`	1.89	0.10	0.53
`CloudComputing`	1.81	0.11	0.51
`GameTheory`	0.60	0.42	0.43
`Bluetooth`	0.45	0.50	0.26
`DVB-T`	0.31	0.42	0.24

Table 4.6: Values of *SEIR model* parameters for the analysis based on *document title*.

Scientific topic	$I(0)$	$E(0)$
`LTE-advanced`	$1.00 \cdot 10^{-5}$	$2.97 \cdot 10^{-3}$
`CloudComputing`	$6.20 \cdot 10^{-4}$	$0.17 \cdot 10^{-3}$
`GameTheory`	$0.51 \cdot 10^{-2}$	$0.96 \cdot 10^{-2}$
`Bluetooth`	$0.19 \cdot 10^{-2}$	$1.77 \cdot 10^{-2}$
`DVB-T`	$0.88 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$

Table 4.7: Values of *SEIR model* parameters for the analysis based on *document title* (initial conditions).

systems. As a consequence, this idea has spread even before the devices using LTE have been commercialized. Instead, looking at the value of σ for `DVB-T` and `Bluetooth`, it can be stated that the *latent period* has a lower value (around 2 months); this might be due to the fact that the technologies on which they are based on are very specific and in the community these ideas began to spread only after being commercialized. Finally, also `GameTheory` shows a small value for the *latent period*, and this again could be due to its interdisciplinarity that allows to already have a developed literature in scientific fields different from the engineering one.

Tables 4.8 and 4.9 and Figures 4.32-4.36 show the results obtained counting papers by *author keywords*. As well as the *SIR model*, also in this case, compared with the previously discussed *document title*, the dynamics are similar for the two different counts and the values of the parameters are almost the same, despite the little changes that could be observed. Thus, the discussion just made is still valid, and the results are confirmed to be meaningful. Moreover, regarding the comparison with the related results found by the *SIR model*, this analysis confirms the improvements given by the consideration of the E class. Once again, the *SEIR model* could be sufficient to describe in a quantitative way the development for `LTE-advanced` and `CloudComputing`, and, at the same time, these results underline that for `GameTheory`, `Bluetooth` and in particular `DVB-T` a more complex analysis is needed.

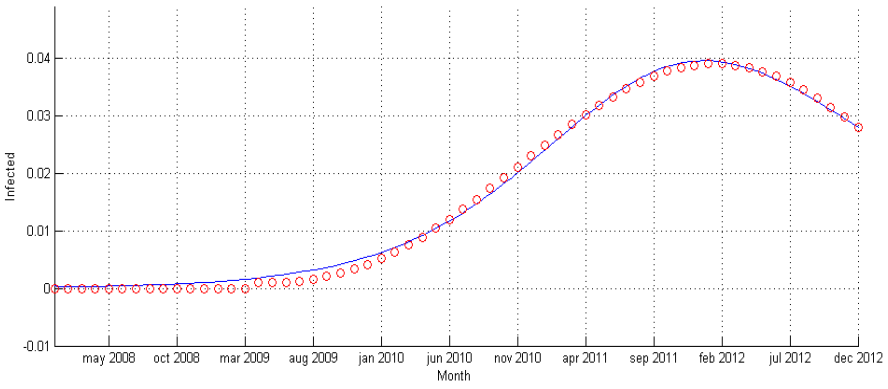

Figure 4.32: Comparison with the theoretical model for `LTE-advanced`.

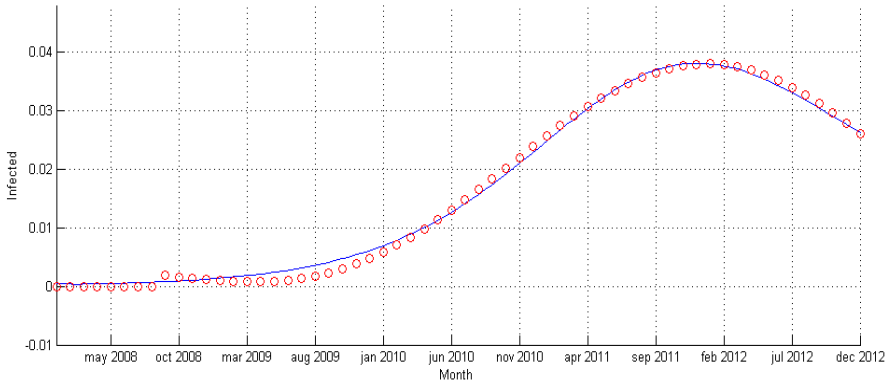

Figure 4.33: Comparison with the theoretical model for `CloudComputing`.

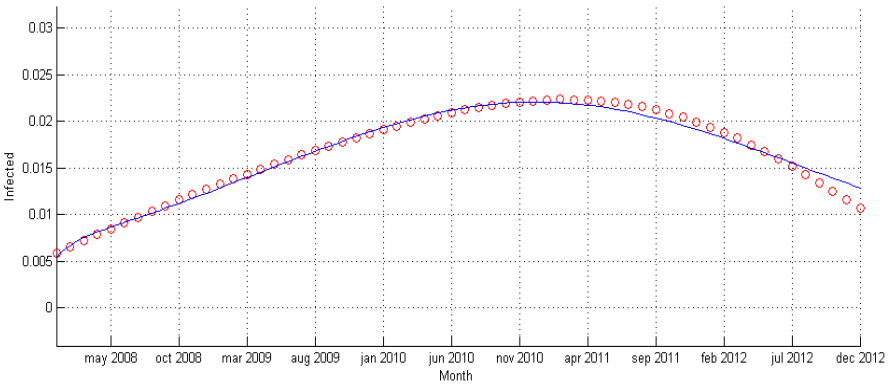

Figure 4.34: Comparison with the theoretical model for `GameTheory`.

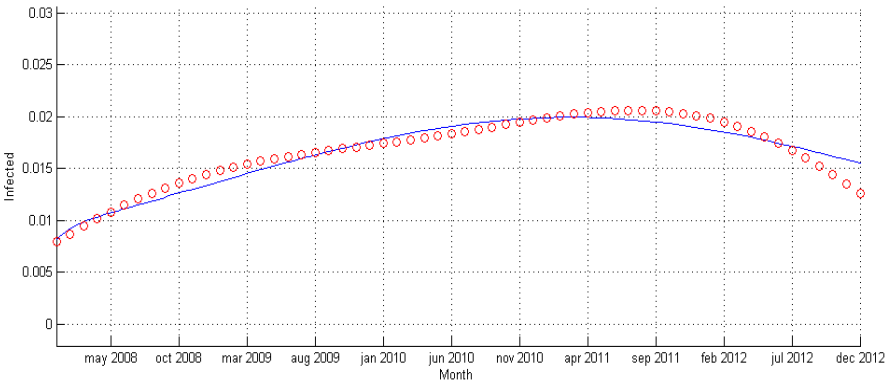

Figure 4.35: Comparison with the theoretical model for `Bluetooth`.

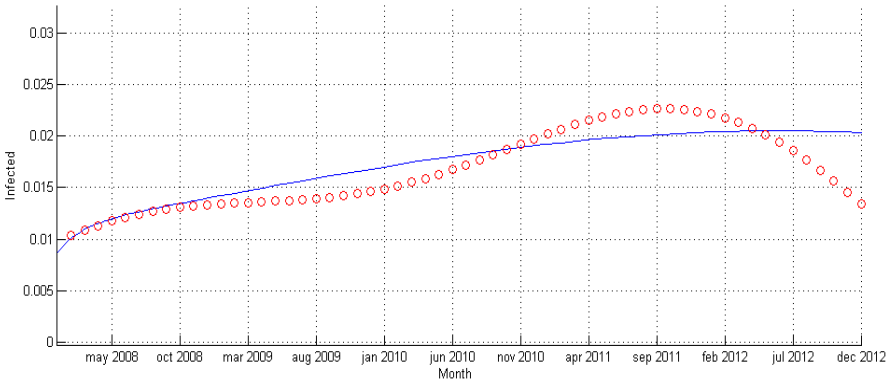

Figure 4.36: Comparison with the theoretical model for `DVB-T`.

Scientific topic	β	σ	γ
`LTE-advanced`	1.97	0.11	0.55
`CloudComputing`	1.97	0.11	0.56
`GameTheory`	0.57	0.31	0.39
`Bluetooth`	0.33	0.44	0.25
`DVB-T`	0.20	0.37	0.15

Table 4.8: Values of *SEIR model* parameters for the analysis based on *author keywords*.

Scientific topic	$I(0)$	$E(0)$
`LTE-advanced`	$6.20 \cdot 10^{-4}$	$1.70 \cdot 10^{-4}$
`CloudComputing`	$6.20 \cdot 10^{-4}$	$5.30 \cdot 10^{-4}$
`GameTheory`	$3.20 \cdot 10^{-3}$	$1.37 \cdot 10^{-2}$
`Bluetooth`	$6.00 \cdot 10^{-3}$	$9.60 \cdot 10^{-3}$
`DVB-T`	$6.60 \cdot 10^{-3}$	$1.00 \cdot 10^{-2}$

Table 4.9: Values of *SEIR model* parameters for the analysis based on *author keywords* (initial conditions).

4.4 SIR model results considering demography

This last point of the discussion concerns the comparison between the results obtained with the basic *SIR model* without and with demography. The main results are summarized in Table 4.10 and Figures 4.37-4.38 for the count considering *document title*, and in Table 4.11 and Figures 4.39-4.40 for the count with *author keywords*. The scientific topics for which the basic *SIR model* gives a good qualitative description of the data have been chosen in order to both better observe the improvements that the consideration of the demography carries and also, if possible, use such a result to infer quantitative insights about the number of articles published per month.

Scientific topic	β	γ	μ	$I(0)$
`LTE-advanced`	0.65	0.46	0.04	$0.30 \cdot 10^{-3}$
`CloudComputing`	0.73	0.50	0.06	$0.30 \cdot 10^{-3}$

Table 4.10: Values of *SIR model* parameters for the analysis based on *document title*.

First of all, observing Table 4.10 and comparing with Table 4.3, it can be stated that when demography is included the *transmission rate* increases slightly for `LTE-advanced` and in a more relevant way for `CloudComputing`, and at the same time the *infectious period* decreases in both case more or less in equal measure. Instead, observing the following figures, in which the dotted red line represent the experimental filtered data, the blue line identifies the development obtained with the model with demography and, finally, the remaining line represents the results given by the model without demography, it can be noted that considering the demography gives a better description of the data trend. Moreover, the values of μ allow to state that, considering the total number of publications counted based on *document title* for each topic, `CloudComputing` has a monthly rate of published articles that is around six times the monthly rate obtained for `LTE-advanced`.

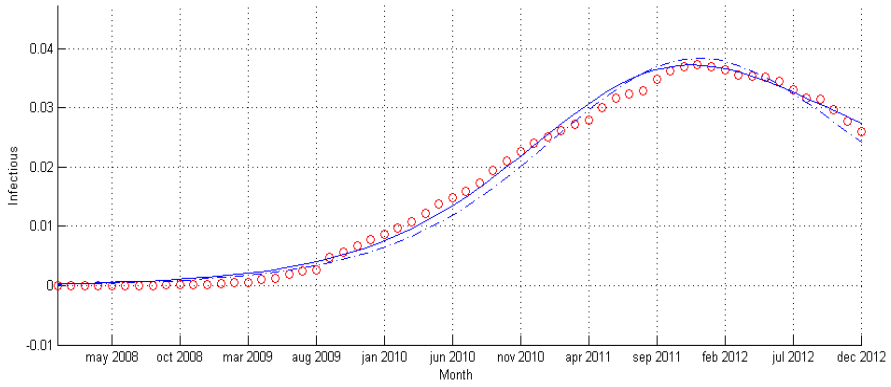

Figure 4.37: Comparison with the theoretical model for `LTE-advanced`.

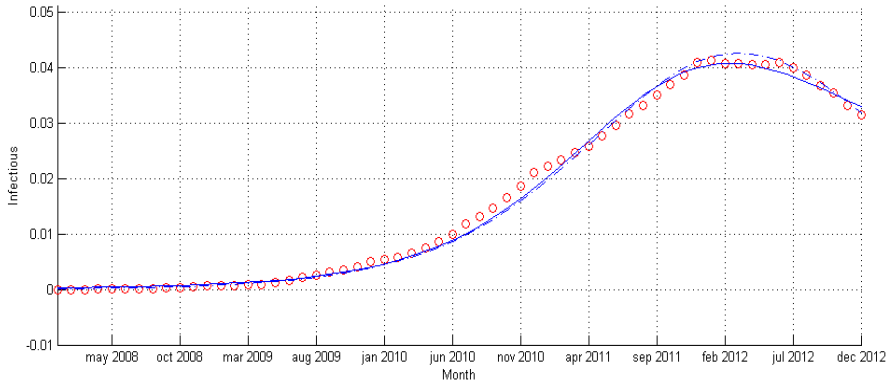

Figure 4.38: Comparison with the theoretical model for `CloudComputing`.

The same considerations can be made for the results obtained considering the *author keywords*, in particular the only differences are that β and γ both increase in equal measure.

Scientific topic	β	γ	μ	$I(0)$
`LTE-advanced`	0.66	0.46	0.04	$0.23 \cdot 10^{-3}$
`CloudComputing`	0.67	0.47	0.04	$0.24 \cdot 10^{-3}$

Table 4.11: Values of *SIR model* parameters for the analysis based on *author keywords*.

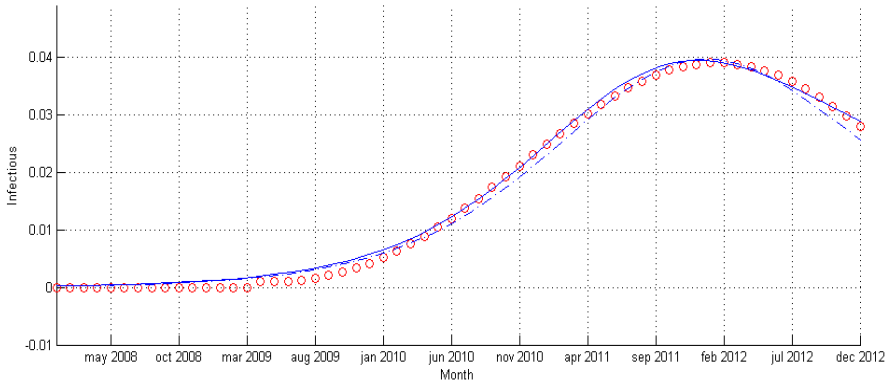

Figure 4.39: Comparison with the theoretical model for `LTE-advanced`.

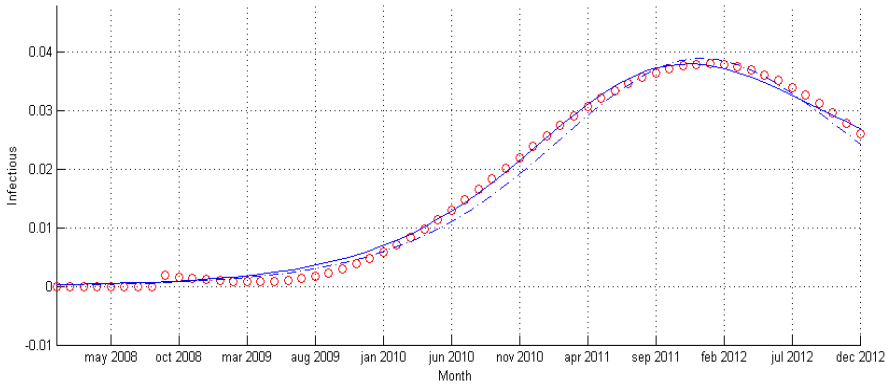

Figure 4.40: Comparison with the theoretical model for `CloudComputing`.

Chapter 5

Conclusions and future works

In this thesis, we considered the application of the epidemiological *SIR model* and some of its extended versions on the diffusion of several scientific research ideas in the worldwide research community. The purpose is to describe the development of an idea in the societal context of the scientific community, thus giving some structured meaning to the simple count of conference articles over time. To this end, we advocated the application of epidemic models for a theoretical comparison that may suggest very useful practical implications.

First, we investigated all the development of the concepts on which epidemics models and bibliometric mechanisms are based on. Then, after highlighting the main aspects of the models used and describing the experimental data, the results have been discussed and analyzed. We focused on 8 different scientific subjects: `BigData`, `SoftwareDefinedNetworking`, `InternetOfThings`, `LTE-advanced`, `CloudComputing`, `GameTheory`, `DVB-T`, and `Bluetooth`. We have found that some of the topics considered are very epidemic; indeed, the application of the basic *SIR model* is sufficient to have a quantitative description of their spread dynamics and, thus, it allows the prediction of the future developments. In particular, among them, `BigData` and `SoftwareDefinedNetworking` are also very contagious, but they are niche topics and are related to a high *transmission rate* in a small *infectious period*; their values of R_0, the *basic reproductive ratio*, are among the largest.

On the other hand, for `CloudComputing` and `LTE-advanced`, which have also raised considerable interest and therefore have a high contagion rate, the simple *SIR model* without demography is not providing a fully accurate description of the phenomenon. Indeed, the comparison between the experimental data and the theoretical curves may exhibit some mismatch. This is due to the fact that these scientific topics are based on some general technologies that must be first studied, analyzed and understood. To take into account of this, we suggested to consider a *latent period*, to evaluate the exposed period of those technologies. In particular, observing the dynamics obtained with the *SEIR model*, it can be stated that this model gives a more precise quantitative description.

Furthermore, we have found that demography can be added to the *SIR model* to improve the results. In this way, it is possible to use the results obtained to infer quantitative insights about the number of articles published. We have found that for `CloudComputing` are approximately published a number of monthly articles that are six times the number related to `LTE-advanced`.

Other topics are found to be less contagious (notably, `GameTheory`, `DVB-T`, and `Bluetooth`). Generally speaking, these topics have either a strong technological footprint or an interdisciplinary character that may slow down their epidemic spreading. Moreover, for these topics we found out that neither the *SIR model* nor its extensions, e.g., the *SEIR model*, are able to achieve a completely acceptable characterization. While the basic *SIR model* is not fully adequate to describe them for the reasons mentioned before, the extended models (in particular, introducing an exposed class) improve the results but do not achieve a fully satisfactory match. For this reason, more complex models may be thought of, e.g., including multiple intermediate states, which leads to a higher-order memory in the dynamics.

As a future development, we may consider transmissions of diseases through multiple vectors or with different populations of carriers, as is the case for many human epidemics. This may be reasonably apply to multidisciplinary topics where the "contagion" spans across different scientific communities.

More in general, the application of epidemic models to the diffusion of

ideas is a promising research field, and can lead to interesting results. A systematic methodology can be applied, for example leading to the development of specific models. At the same time, it must be considered that idea spreading is not just a matter of presentation. The diffusion of science is also based on the concrete value and solidity of the theory behind it. For this reason, a better understanding of this phenomenon can also be reached through the superimposition of different processes, combined from evaluative and critical observations from philosophy of science. In this sense, an advancement of our study can also lead to a cross-fertilization between apparently distant disciplines, and possibly new proposals, for both social sciences and scientific sociology.

References

[1] J. M. Cattell, "Statistics of American psychologists," American Journal of Psychology, vol. 14, no. 3/4, pp. 310-328, Jul.-Oct. 1903.

[2] J. M. Cattell, "The advance of psychology," Science, Vol. 8, pp. 533-541, Oct. 1898.

[3] B. G. Miner, "The changing attitude of American universities toward psychology," Science, vol. 20, no. 505, pp. 299-307, Sep. 1904.

[4] E. F. Buchner, "Ten years of American psychology," Science, vol. 18, no. 451, pp. 233-241, Aug. 1903.

[5] E. F. Buchner, "A quarter century of psychology in America: 1878-1903," American Journal of Psychology, vol. 14, pp. 402-416, Jul.-Oct. 1903.

[6] C. A. Ruckmich, "The history and status of psychology in the United States," American Journal of Psychology, vo. 23, no. 4, pp. 517-531, Oct. 1912.

[7] C. A. Ruckmich, "The last decade of psychology in review," Psychological Bulletin, vol. 13, no. 3, pp. 109-120, Mar. 1916.

[8] S. W. Fernberg, "The American psychological association: a historical summary, 1892-1930," Psychological Bullettin, vol. 29, no. 1, pp. 1-89, Jan. 1932.

[9] S. W. Fernberg, "On the number of articles of psychological interest published in the different languages," American Journal of Psychology, vol. 28, no. 1, pp. 141-150, Jan. 1917.

[10] S. W. Fernberg, "Publications, politics and economics," Psychological

Bulletin, vol. 35, no. 2, pp.84-90, Feb. 1938.

[11] W. Denis, "Bibliographies of eminent scientists," Science, vol. 79, no. 3, pp. 180-183, Sep. 1954.

[12] W. Denis, "Age and productivity among scientists," Science, vol. 123, no. 3200, pp. 724-725, Apr. 1956.

[13] F. J. Coles and N. B. Eales, "The history of comparative anatomy, Part 1: a statistical analysis of the literature," Science Progress, vol. 11, no. 44, pp. 578-596, Apr. 1917.

[14] E. W. Hulme, "Statistical bibliography in relation to the growth of modern civilization," London Grafton, 1923.

[15] P. L. K. Gross, E. M. Gross, "College libraries and chemical education," Science, vol. 66, no. 1713, pp. 385-389, Oct. 1927.

[16] P. Otlet, "Traité de documentation: le livre sur le livre, théorie et pratique," 1934.

[17] A. Pritchard, "Statistical bibliography or bibliometrics?," Journal of Documentation, vol. 25, no. 4, pp. 348-349, 1969.

[18] A. Pritchard, "Bibliometrics and information transfer," Research in Librarianship, vol. 4, pp. 37-46, 1972.

[19] W. G. Potter, "Introduction to bibliometrics," Library Trends, vol. 30, pp. 3-7, 1981.

[20] E. Garfield, "Citation indexes for science: a new dimension in documentation through association of ideas," Science, vol. 122, no. 3159, pp. 108-111, Jul. 1955.

[21] F. Narin, "Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activit,", Computer Horizons, 1976.

[22] A. J. Lotka, "The frequency distribution of scientific productivity," Journal of the Washington Academy of Science, vol. 16, no. 12, pp. 317-323, 1926.

[23] S. C. Bradford, "Documentation," Whashington, Public Affairs Press,

1950.

[24] G. K. Zipf, "Psycho-biology of language: an introduction to dinamic philology," London, Routledge, 1935.

[25] N. De Bellis, "Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics," Lanham, MD, The Scarecrow Press, 2009.

[26] ISI Web of Science database, available online at: <http://wokinfo.com/products_tools/multidisciplinary/webofscience/>.

[27] Scopus database, available online at: <http://www.info.sciverse.com/scopus/scopus-in-detail/facts>.

[28] Google Scholar database, available online at: <http://scholar.google.it/>.

[29] D. Bernoulli, "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir," Mém Math Phys Acad Roy Sci, Paris 1766.

[30] W. Hamer, "Epidemic diseases in England: the evidence of variability and of persistency of types," The Lancet, vol. 167, no. 4305, pp. 569-574, Mar. 1906.

[31] R. Ross, "The prevention of malaria," New York, E.P. Dutton & company, 1910.

[32] W. O. Kermack, A. G. McKendrick, "A contribution to the mathematical theory of epidemics," Proc. R. Soc. Lond. A, vol. 115, no. 772, pp. 700-721, 1927.

[33] H. Abbey, "An examination of the Reed-Frost theory of epidemics," Hum Biol., vol. 24, no.3, Sep. 1952.

[34] R. M. Anderson, R. M. May, "Infectious diseases of humans," Oxford Science Publications, 1991.

[35] R. M. Anderson, R. M. May, "Regulation and stability of host parasite population interaction," Journal of Animal Ecology, vol. 47, no. 1, pp. 219-247, Feb. 1978.

[36] R. M. Anderson, R. M. May, "Population biology of infectious diseases,"

Nature 280, pp. 361-367, Aug. 1979.

[37] R. M. Anderson, R. M. May, "The population dynamics of microparasites and their invertebrate hosts," Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, vol. 291, no. 1054, pp. 451-524, Apr. 1981.

[38] P. D. O'Neil, D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison, "Analysis of infectious disease data from household outbreaks by Markov chain Monte Carlo methods," Journal of the Royal Statistical Society. Series C, vol. 49, no. 4, pp. 517-542, 2000.

[39] P. D. O'Neil, G. O. Roberts, "Bayesian inference for partially observed stochastic epidemics," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 162, no. 1, pp. 121?129, 1999.

[40] G. J. Gibson, E. Renshaw, "Estimating parameters in stochastic compartmental models using Markov chain methods," Oxford JournalsScience: Mathematics Mathematical Medicine and Biology, vol. 15, no. 1, pp. 19-40, 1998.

[41] C. Castillo-Chavez, "Mathematical and statistical approaches to AIDS epidemiology," Lecture Notes in Biomathematics, Springer-Verlag New York, vol. 83, 1989.

[42] H. W. Hethcote and J. W. Van Ark, "Modeling HIV transmission and AIDS in the United States," Lecture notes in Biomathematics 95, Springer-Verlag, Berlin, Heideiberg, New York, 1992.

[43] H. W. Hethcote and J. A. Yorke, "Gonorrhea transmission dynamics and control," Springer, New York, 1984.

[44] M. J. Keeling, P. Rohani, "Modeling infectious diseases," Princeton University Press, New Jersey, USA, 2008.

[45] M. Greenwood, A. Bradford Hill, W. W. C. Topley, J. Wilson, "Experimental epidemiology," Journal of the Royal Statistical Society, vol. 100, no. 1, pp. 103-106, 1936.

[46] M. S. Bartlett, "Deterministic and stochastic models for recurrent epi-

demic," Proc. Third Berkeley Symp. on Math. Statist. and Prob., vol. 4 (Univ. of Calif. Press), pp. 81-109, 1956.

[47] A. Lajmanovich and J. A. Yorke, "A deterministic model for gonorrhea in a non homogeneous population," Mathematical Biosciences, vol. 28, no. 3-4, pp. 221-236, 1976.

[48] K. Cooke and J. Yorke, "Some equations modeling growth processes and gonorrhea epidemics," Mathematical Biosciences, vol. 16, no. 1-2, pp. 75-101, Feb. 1973.

[49] S. Leinhardt, "Social networks: a developing paradigm," American Anthropologist, vol. 80, no. 3, pp. 686-688, Sep. 1978.

[50] J. Scott, "Social network analysis: a handbook," SAGE Publications Ltd, 2000.

[51] F. Harary, "Graph theory," Reading, MA: Addison-Wesley, 1969.

[52] B. Ballobas, "Graph theory," New York, Springer, 1979.

[53] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-word networks," Nature,393, pp. 440-442, 1998.

[54] C. Moore and M. Newman, "Epidemics and percolation in small-word networks," Phys. Rev. E 61, 5678-5682, 2000.

[55] R. Albert, H. Jeong, A. Barabasi, "Diameter of the world-wide web," Nature 401, 130-131, Sep. 1999.

[56] R. Pastor-Satorras, A. Vespignani, "Epidemic spreading in scale-free networks," Phys. Rev. Lett. 86, 3200, Apr. 2001.

[57] A. F. Rozenfeld, R. Cohen, D. Avraham, S. Havlin, "Scale-free networks on lattice," Phys. Rev. Lett. 86, 218701, 2002.

[58] D. J. Preshiya, C. D. Suriyakala, "Immunized SIR model for routing in sparse mobile ad hoc networks," Proceedings of 2013 IEEE Conference on Information and Communication Technologies, pp. 874-878, 2013.

[59] A. Jafarabadi, M. A. Azgomi, "An SIR model for the propagation of topology-aware active worms considering the join and leave of hosts," 2011

7th International Conference on Information Assurance and Security (IAS), IEEE, 2011.

[60] A. Jafarabadi, M. Abdollahi Azgomi, "On the impacts of join and leave on the propagation ratio of topology-aware active worms," Proc. 4th International Conference on Security of Information and Networks (SIN'11), Sydney, Nov. 2011.

[61] A. J. Ijspeert, T. Masuzawa, S. Kusumoto, "Biologically inspired approaches to advanced information technology," Lecture Notes in Computer Science, vol. 3853, 2006.

[62] D. F. Bernardes, M. Latapy, F. Tarissan, "Relevance of SIR model for real-world spreading phenomena: experiments on a large-scale peer-to-peer system," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), At Istanbul, Turkey, 2012.

[63] K. Zhu and L. Ying, "Information source detection in the SIR model: a sample path based approach," in ITA?13: Proc. of 2013 Inform. Theory and Applications Workshop, 2013.

[64] Z. Chen, K. Zhu, L. Ying, "Detecting multiple information sources in networks under the SIR model," Information Sciences and Systems (CISS), 2014 48th Annual Conference.

[65] S. Tang and B. Mark, "Analysis of virus spread in wireless sensor networks: an epidemic model," In Design of Reliable Communication Networks, pp. 86-91, Washington, USA, Oct. 2009.

[66] J. E. Hirsch, "An index to quantify an individual?s scientific research output," Proc. of the National Academy of Sciences of the United States of America, vol. 102, no. 46, 2005.

[67] H. P. F. Peters, A. F. J. Van Raan, "On deterministic of citation scores-a case of study in chemical engineering," Journal for the American Society for Infor. Science, vol. 45, no. f, pp. 39-49, 1994.

[68] S. Baldi, "Normative versus social constructivist processes in the collocation of citations: a network-analytic model," American Sociological Review,

vol. 63, no. 6, pp. 829-846, 1998.

[69] H. M. Collins, "Tantalus and the alieus: publications, audiences and the search for gravitational waves," Social studies of Science, vol. 29, no. 2, pp. 163-197, 1999.

[70] W. Goffman, V. A. Newiji, "Generalization of epidemics theory: an application to the transmission of ideas," Naiure 204, pp. 225-228, 1964.

[71] W. Goffman, "Mathematical approach to the spread of scientific ideas?the history of mast cell research," Nature 212, pp. 449-452, 1966.

[72] L. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, C. Castillo-Cháve, "The power of a good idea: quantitative modeling of the spread of ideas from epidemiological models," Physica A 364, pp. 513?536, 2006.

[73] I. Z. Kiss, M. Broom, P. Craze, I. Rafols, "Can epidemic models describe the diffusion of topics across disciplines?," Journal of Informetrics 4(1), pp. 74-82, 2010.

[74] S. M. Blower, P. M. Small, R. C. Hopewell, "Control strategies for tubercolusis epidemics: new models for old problems," Science, vol. 273, no. 5274, pp. 497-500, Jul. 1996.

[75] IEEEXplore engine, online database at the Institute of Electrical and Electronic Engineers, available online at:
<http://ieeexplore.ieee.org/Xplore/home.jsp>.

[76] B. Godin, "On the origins of bibliometrics," Scientometrics, vol. 68, issue 1, pp. 109-133, Jul. 2006.

[77] J. Travers and S. Milgram, "An experimental study of the small world problem," Sociometry, vol. 32, no. 4, pp. 425?443, 1969.

[78] D. J.Watts and S.H. Strogatz, "Collective dynamics of ?smallworld? networks," Nature, vol. 393, no. 6684, pp. 440?442, 1998.

[79] R. Albert, and A. Barabasi, "Statistical Mechanics of Complex Networks," Reviews of Modern Physics 74, 47, 2002.

[80] ACM digital library, online database at the Association of Computing

Machinery, available online at: <http://dl.acm.org>.

[81] L. Subelj, M. Bajec, "Model of complex networks based on citation dynamics," Proceedings. of the WWW Workshop on Large Scale Network Analysis 2013 (LSNA '13), pp. 527-530.

[82] Z.-X. Wu, P. Holme, "Modeling scientific-citation patterns and other triangle-rich acyclic networks," Phys. Rev. E, 80(3):037101, 2009.

[83] K. E. Nelson, C. F. Williams, "Early history of infectious disease," Jones and Bartlett Publishers.

[84] Francis Aidan Gasquet, "The Black Death of 1348 and 1349," 2nd ed. London, England: George Bell and Sons, 1908.

[85] P. Salway, W. Dell, "Plague at Athens," Greece & Rome, Second Series, vol. 2, no. 2, pp. 62-70, Jun. 1955.

[86] W. R. Gilks, "Markov Chain Monte Carlo," Wiley Online Library, 2005.

[87] M. J. Keeling, K. T. D. Eames, "Networks and epidemic models,"J. R. Soc. Interface, vol. 2, pp. 295?307, Jun. 2005.