KNOWLEDGE NETWORKS: THE CASE OF WIKIPEDIA

Vinko Zlatić
Centro SMC CNR-INFM, Dipartimento di Fisica, University of Rome "Sapienza", Piazzale Aldo Moro 5 00185 Rome, Italy
Theoretical Physics Division, Rudjer Boskovic Institute, P.O.Box 180, HR-10002 Zagreb, Croatia

Keywords: Expert Networks, WWW, Wikipedia graphs

Contents

Summary

In this chapter we introduce some different definitions for the wide class of knowledge networks. They can be intended as a web of expert in a certain field, as a web of connections between pieces of research or more generally as the network in which concepts are related to each other. A paramount example of the last kind of network is given by the online encyclopedia Wikipedia to which a specific analysis is devoted.

1. Introduction

Under the name of "knowledge networks" we can describe a variety of different systems according to the particular meaning considered. This is related to the fact that also network theory is used in many different fields of science where experts use very different language, have different objectives and focus on different subjects. Given these premises ``knowledge networks" represent one of the best way to proceed to a disambiguation. Indeed, social scientists use this term in a completely different context from psychologists or cognitive scientists. The knowledge networks include citation-network of references in scientific papers, network representation of the World-Wide Web with the educational or scientific information, network databases (for example, protein interactions), network of different experts, etc. Whenever tempting a general overview of such a distinct type of networks we have to consider that all of them in their respective fields are considered as knowledge networks.

More generally in the field of knowledge networks we include the "maps" of collaborating institutions and/or experts, used for evaluating and creating scientific and technological policies. Maps of citations between different scientific papers, maps of available data, educational material or library content are all used to allow an easier html navigation (in the search for additional learning resources or for the evaluation of the authors of these materials). These maps are very important since this representation of interconnected concepts are pieces of information used for better understanding of the ways in which person is learning or organizing its memory. Sometimes it is hard to distinguish in which category a real network belongs. For example in the network of mutual citations of scientific papers one can choose to include non-scientific references to the pages like Wolfram's Research or Wikipedia (http://www.Wikipedia.org). What one can certainly say is that most knowledge networks describe certain available paths to new knowledge. That is obtained through connections between experts in some field, or through citation of important papers or through semantic interdependence of some concepts.

In this chapter I will give a short introduction focused on the first two types of knowledge networks and then I will more focus on concept networks with particular emphasis on the case of study of the Wikipedia network.

2. Expert Networks

Expert networks are social communities of people with different expertises. In the old days a talented person was able to work as a scientist, architect, painter, writer, as Leonardo Da Vinci did. Nowadays with the growth of the size of the available knowledge growth, people tend to be more specialized. Even in science some people were trained as engineers, some as physicists, some other as chemists. In a modern society enormous amount of knowledge that was produced in the last centuries pushes us to be more and more specialized in our training. For example in physics the PhD students specialize in some more specific area of physics like statistical mechanics or optics or particle physics and then proceed their career in an even more narrow subjects like numerically computed density function approximations or complex networks or Raman spectroscopy. Also in the technology we are witnessing a similar phenomenon; our society is producing more and more complex goods and services. While telephone was an invention of at most two persons (non collaborating) Antonio Meucci and Graham Bell the present situation is very different. Now in order to produce a new model of mobile phones we have teams of microprocessor developers, designers, software developers and many others. If we want to produce any successful product, we need to bring together several experts from very diverse fields and make them work together. In this way the collaboration of these experts form an expert network. Another interesting aspect of this expert networks is the validation of performance in complex business environments. Companies or institutions would like to know patterns of interactions among their employees and to see how some people affect quality of communications between different working groups or experts [Hildreth 2005]. In the field of social sciences there are numerous experts using network theory in their job as business consultants. Their goal is to optimize business processes in such a way that employees from different working groups most easily exchange their knowledge for the benefit of their employer. A third application of expert network usage is in the development of regions or start up of companies. To access the potential competitive advantage of certain geographical location, scientists provide details of local expert networks to policy makers or investors. Dense and diverse network of experts and specialized companies [Cesar 2009] can help underdeveloped regions or new companies in their development.

2.1. Citation Networks

As noted before we live today in a complex world in which it is getting harder and harder to follow new theoretical advances. While just 50 years ago a young person could have learned important things about its job from his grandfather, these days even our fathers can not really understand new high tech jobs. New phrases like "live long education", "knowledge based economy", "working force flexibility", etc. have entered in the language of every modern politician. Indeed, new technologies like web 2.0 present opportunity for people to learn new things, enforce their old knowledge, create their own curriculums and so on. The vast majority of educational material, social networks for exchange of knowledge, free access textbooks, tutorials, etc are well hidden in the enormity of contemporary internet. Although knowledge is freely available and in different ways indexed in many research engines it is hard to find good and well prepared resources. What is true for the non-academic acquisition of knowledge holds also for the academic one. The contemporary world of science is crowded by new journals and new sources of information. The scientists of the fifties in the last century had a possibility to read the entire content of *Physical Review* on a a weekly basis. The current overproduction of ``knowledge" discourages scientists to follow even all the relevant papers in their own small field of research. Struggling in the sea of information that is scattered in the books, journals and World Wide Web has become so complicated that a whole issue of *Proceedings of the National Academy of Sciences (USA)* was completely devoted to the so-called mapping of knowledge. In order to find efficiently a piece of research we need, we have to understand how new knowledge emerges and disappear and the old one becomes relevant again (for example papers not cited for a long period may become one day modern and relevant). These issues are now an interdisciplinary subject studied by a variety of scientists.

2.2. Networks of Citations between Scientific Papers

There is a number of studies focussed in the description of citation networks among scientific papers. The papers published on a peer reviewed journals and stored in a database represent the nodes and the citations found in these papers are the directed links among the nodes in the network. What is the reason why such networks are interesting to scientists? There are at least some different causes of interest. Firstly, this is a typical problem for the bibliometrics community. Bibliometrics experts are trying to find ways to evaluate scientific output of given scientist, research group, university or the whole country based on their publication history. What policy makers and funding institutions would like to have is quantitative information on the expected quality of the research they are going to fund. Number of citations and functions like the *h-index* are probably the most crucial parameters to assess the quality of the people involved in some piece of research. More accurate analysis is made by using independent peer reviews of scientists, personal interviews and other bibliometric methods. It is hard to expect that subjective biases can be completely deleted in the peer reviews and personal interviews because they could accidentally call a friend of a scientist for its reviewer in a case or could fall to the charm of scientist interviewed. Therefore the bibliometers are trying to find nonsubjective numerical criteria to access value of the human force in the projects and they extensively use the tools and ideas from network theory in order to do that. Complementary to this, policy makers are also trying to understand which parts of science are getting more ``hot" and to direct the extra funds to the growing fields. Bibliometric informations obtained by usage of network theory can greatly help in evaluation of such fields. Secondly, there is a strong interest from social scientist interested in habits and working ethics of different scientific communities. They would like to know and understand how the force of authority is delegated in some research communities, how the scientists choose to cite their papers and so on. Thirdly, there is a growing interest in the automatic data and knowledge management. More and more journals and independent sites are trying to use network theory to help their readers in the navigation through the vast field of scientific papers. Their aim is to find papers which based on the content analyzed through pattern recognition programs and on their position in the citation network resemble most the papers which user actively reads. Then they can recommend those papers to the reader and actually help him in the research, saving valuable time.

2.3. Semantic Networks

Maybe the most precise or at least the most natural definition of knowledge network is network which is drawn from interdependence of different types of knowledge. This type of networks is called *semantic networks*, because nodes represent concepts and links represent their semantic relationship [Sowa (2000)]. Take for example a formula for the roots of the second order equation $ax^2 + bx + c = 0$

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Anyone who wants to learn this formula first has to know how to calculate root and square, how to divide and multiply and how to add and subtract. A non expert also needs to know what an equation is and what the roots of a second order equation are. In order to organize and use new knowledge we have to rely and connect it to the knowledge we have already once acquired. Studies of such connections can help in devicing new ways to organize knowledge. They can be used in more sofisticated versions of artificial inteligence and so on. Semantic networks are actualy an old tool developed in the 50's of the last century to enable programmers to create a common language with which they can represent natural languages in the machine framework.

Psychologists since a number of years have used *associative networks* where nodes represent concepts in memory and links represent associations between those concepts. For example they test a number of people in order to make associations between some given set of words [Steyvers et al. (2005)]. Most healthy individuals have similar associative patterns and those networks look similar. Such networks are related to the way we organize our memory (especially a short term memory), see for example [Glautier et al (1999), Paulsen et al (1996)] and are also used to validate impact of different types of brain condition (head traumas, psychological illnesses) on a patient memory. Although there is a lack of clear connection between word connected just by associations and some kind of formal knowledge, this type of networks clearly fall in the category of semantic networks [Capocci et al. 2005]. In computer science the same term - associative networks is reserved for the semantic network of objects in some computer languages.

3. The WWW and Internet as Knowledge Network

Today we are making strong advances in the mapping of available knowledge that exists on the WWW. Just 10 years ago the situation was completely different. Search for given material was performed only by knowing the exact address (URL) of the content needed or by random browsing of domains where a researcher was expecting to find the desired topics , or through portals where the information was stored in hierarchical way (Yahoo!). Most of the reading was still performed in the old university or public libraries. At the time some of the programmers were already using *crawlers* in order to find the interesting content on the web. Crawlers are small programs which travel from web page to web page following hyperlinks either completely random, or following some preconditioned pattern. These crawlers would also search for given pieces of text that the programmer would provide before his search. With such automation the searches for additional knowledge were easier and faster. Nevertheless the WWW kept growing all the time at an exponential rate and as a result the crawlers need more and more time to find a desired content. The obvious solution was using crawlers to pick up as much as possible information about the pages that they have visited and store them somewhere so that the search would not start from zero every time when it is performed. Analysing this data was also a huge task. Some queries would give millions and millions of pages with required regular expression, but only a small portion of them would really be useful to the searcher. It was of great importance to develop algorithms that will somehow filter and rank the content of the pages so that this scattered knowledge finds its way to the end user. The best example of such algorithm (but not the only one) is *PageRank*, the algorithm which is behind huge success of the Google search engine.

Page Rank is an algorithm, which simulates a random crawl through the real WWW. It counts how often a crawler visits a certain page in an eternal random walk. The frequency of visits is given as a number attached to the page. This number is assumed to represent the overall importance of the page in the WWW. Page Rank algorithm works very well in the categorization of pages because through random walk statistics it integrates the information on the connectivity pattern of the whole WWW and not just about local connectivity or any other measure. Naturally, some pages have very high Page Rank but can also have some words which are of very low importance for that page. For example a researcher which is interested in impacts of climate change on the Arctic could use search words ``Impact of climate change" or ``Polar ice rim". This researcher then expects to find a number of science related pages and not the list of songs of some band. Therefore the additional heuristics which tries to understand real needs of every researcher and the content of the page had to be developed. In this case developers often use *bipartite* and *weighted* networks to represent the data. These days the WWW is also full of so called *spam pages*. These pages are developed for marketing purposes with the aim of enhancing the rank of specific pages. They do not have any useful content and their whole structure is developed with the only reason to increase its Google Page Rank. Most of the companies pay advertising by the number of visits that some page gets, and being on an higher position in the Google ranking ensures more visitors to the spam pages, which in turn generates advertising money for the owners. Big companies like Google or Yahoo! on the other hand spend a large amount of money in order to fight those spam pages because in general they decrease the quality of the answer to a specific query. As we have already seen, the usage of network theory is essential in searching knowledge around the WWW. With the new technological advances the network theory arises as an inevitable tool. A simpler solution to look for specific information is the creation of *ad hoc* encyclopedia; this is the case of Wikipedia that we are going to describe in the following section.

4. Wikipedia Analyses

Wikipedia is an encyclopedia whose pages are located in the World Wide Web. It is lead by the idea that anyone who thinks that he or she can contribute to the quality of articles. Everybody can write new articles or edit old articles or add a media material. Everybody is encouraged to connect articles via hyperlinks and engage in discussion with other contributors (in the rest of the text we will call them “Wikipedians”, because it is the name they choose for themselves).

Such an editorial policy and the great popularity of the site has developed in the public a fierce debate about the quality of `` knowledge'' presented in articles. The debate has become particularly infuriated after independent analysts claimed that Wikipedia is the most often consulted WWW-page in several categories. Recent research confirmed that the quality of Wikipedia articles is no second to the quality of the material in traditional encyclopaedias, such as the Encyclopaedia Britannica [Giles (2005)]. The results of this research suggest that, probably, the structure of hyperlink connections between different pages with similar context is well done. In this text we focus on the network of hyperlinks in the Wikipedia.

4.1 Building of Wikipedia Graphs and Ensembles

Crucial point in the Wikipedia evolution is the rules of construction that are applied as a community policy. Firstly, all the articles are produced, in a large number of iterations, until all Wikipedians interested in the topic agree that article quality is good enough. Secondly, there is a precise request to link the structure i.e. the new articles have to point with its hyperlink to at least one older article which already exist in the Wikipedia to ensure appropriate ``connectedness'' of the exposed ``knowledge'' (It exist also the symmetric request to avoid as much as possible the “orphans” that is new pages that are not linked from the old ones). Wikipedia community has evolved from the initial requirement that each article is written from a neutral viewpoint (NPOV: Neutral point of view - policy). Effectively all the possible opposed views related to the article shall be discussed in the discussion pages which exist for each article. In this way every article is an example of cooperation of a number of independent persons and a good representative of ``average point of view". Because of such building policy and the lack of central structures, Wikipedia grows as a complex system self-determined by local stochastic rules. The writers of articles do not have full information on all the existing Wikipedia pages. Therefore they choose to which page they will connect their article on the basis of the belief of the majority of contributors. They decide what is relevant for the contents introduced. It is fair to claim that Wikipedia represents the only real system which grows with the only restrictions given by available information. Each Wikipedia can be seen as a directed network whose vertices correspond to different articles, while directed edges match hyperlinks between them. It is possible to find more than 200 Wikipedias on the World Wide Web, They are written in different languages, with different number of articles, and different structures of hyperlinks that connect them. All the Wikipedias are constantly growing in time with the addition of new articles, and hyperlinks between them.

Although Wikipedias in different languages grow mostly independently one from one another, there is a certain number of Wikipedians, which contribute to writing articles on two or more different languages. Certain number of articles is copied and translated from a one Wikipedia to another. Also the older well established Wikipedias also influence, with their standard of writting and individual rules, the younger once. Nevertheless, it can be claimed that articles in different languages originate independently, and that the growth of any Wikipedia graph can, in principle, be described as an autonomous process. With this plausible assumption of unique stochastic rules of growth and autonomy in the process of construction of each Wikipedia, they represent a real case of statistical ensemble of replicas of the same social system. That is a very important property. Almost any other real complex networks studied and observed is unique. This is the case of the Internet and WWW, or it can have only a small number of available realizations, as for example, protein interactions. It is important, from the epistemological perspective to study how much the hypothesis of a statistical ensemble is really applicable for the description of a social-complex system such as Wikipedia. In its essence Wikipedia represents a network of knowledge. A type of encyclopedical knowledge, edited with non-standard editorial policies, but no less interesting then the network of quotations or a web pages. One of the first steps in any new research is the collection of all possible data relevant for this research. In this sense, the research and of Wikipedia is extremely important for the whole area of a research of knowledge networks.

-
-
-

Bibliography

A.-L. Barabási, R. Albert “Emergence of Scaling in Random Networks” *Science* 286, 509 – 512 (1999) [The first paper on scale-free Networks].

A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, “Detecting communities in large networks Physica A 352, 669–676 (2005) [A paper on the analysis of spectral properties of graphs in order to detect the communities].

A. Capocci, V.D.P. Servedio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, G. Caldarelli “Preferential attachment in the growth of social networks: the case of Wikipedia” Physical Review E, 74 036116, (2006) [A paper on the statistical properties of Wikipedia].

D. Garlaschelli and M. I. Loffredo, “Patterns of Link Reciprocity in Directed Networks” *Physical Review Letters* 93, 188701 (2004) [A paper on the property of reciprocity in a graph]

J. Giles “Special Report Internet encyclopaedias go head to head” *Nature* 438, 900-901 (2005) [A report on the comparison between Wikipedia and Britannica].

S. Glautier, K. Spencer, “Activation of alcohol-related associative networks by recent alcohol consumption and alcohol-related cues”, *Addiction*, {94}, 1033 - 1041 (1999) [A paper on cognitive associations].

P. Hildreth, C. Kimble, “Knowledge Networks: Innovation Through Communities of Practice”, *K-Now International, USA, University of York, UK*, IGI Publishing (2004) [A paper on the knowledge networks].

C. A. Hidalgo, R. Hausmann “The Building Blocks of Economic Complexity”, *Proceedings of the National Academy of Science*, 106, 10570-10575 (2009). [A paper on the development of nations and their trade].

G. Z. López, V. Zlatic, C. Zhou, H. Stefancic, J. Kurths “Reciprocity of networks with degree correlations and arbitrary degree sequences” *Physical Review E)*, 77, 016106 (2008) [An analysis of the correlation between degree and reciprocity].

S. Maslov, K. Sneppen and A. Zaliznyak “Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet”, *Physica A* 333, 529-540 (2004) [A paper on the definition of a null case for the degree correlations].

R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon “Superfamilies of Evolved and Designed Networks” *Science* 303, 1538 – 1542 (2004) [a review on regularities on various networks].

M.E.J. Newman "The structure and function of complex networks". *SIAM Review* 45 167–256 (2003) [A nice review on scale-free networks]

J. S. Paulsen, R. Romero, A. Chan, A. V. Davis, R. K. Heaton, D. V. Jeste, “Impairment of the semantic network in schizophrenia”, *Psychiatry Research*, 63, 109-121 (1996) [A paper on logical associations in patients affected by schizophrenia].

J. F. Sowa, “Knowledge Representation: Logical, Philosophical, and Computational Foundations”, *Brooks Cole Publishing Co*., Pacific Grove, CA, (2000) [a Text on the organization of knowledge].

M. Steyvers, J.B. Tenenbaum “The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth” *Cognitive Science* 29, 41–78 (2005) [The results of a psychological experiment and the network of associations].

V. Zlatić, M. Božičević, H. Štefančić, M. Domazet,"Wikipedias:Collaborative web-based encyclopedias as complex networks", *Physical Review E* 74, 016115 (2006) [An analysis of the properties of various Wikigraphs].

Zlatić V, Štefančić H. Influence of reciprocal edges on degree distribution and degree correlations. *Physical Review E*. 2009;80(1):016117. Available at: http://link.aps.org/doi/10.1103/PhysRevE.80.016117

Biographical Sketch

Vinko Zlatić is born the 13.9.1974 in Zagreb, Croatia and got a PhD in Physics in year 2009 in the University of Zagreb. Most of his scientific activity has been devoted to the topological analysis of the Wikipedia Graph and reciprocity in modeled and real networks.