OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author's version published in:
http://oatao.univ-toulouse.fr/22012

Official URL
DOI : http://doi.org/10.1109/ISKO-Maghreb.2013.6728198

To cite this version: Barros Sampaio, Ricardo and El Haddadi, Anass and Bahsoun, Wahiba and Dousset, Bernard Data on Social Network Analysis on Malaria. (2014) In: 3rd International Symposium ISKO-Maghreb, 8 November 2013 - 9 November 2013 (Marrakech, Morocco).

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr
Abstract - In order to understand the existent and in
formation relationships that might surface from scientific
publications it is necessary to use methods for data and network
analysis, besides the concepts of information flow and knowledge
creation within organizations. The aim of this work was to show
the application of tools and methods for data and network
analysis and how those methods can bring forward proposals to
improve the understanding of a specific disease, Malaria, within
the scientific publications. The study of relational data retrieved
from different databases might be elucidating in an environment,
health system, which happens to be quite complex due to its
many elements involved. How these elements interact within each
other in order to fulfill the need to serve the population needs to
be understood. The study of scientific publications is just one of
the areas that might help decision makers have a better grasp on
how to deal and focus on the needed areas.

Keywords - Scientific and technological vision, bibliometric,
grahps, social networks, geo-strategy, visualization, interactivity,
 neglected diseases, malaria

I. INTRODUCTION

The use of data and social analysis tools has become a very
important technique in order to find and retrieve relevant
information. With the use of massive storage spaces and great
amounts of data new technologies for sorting, retrieving and
visualizing this information is necessary. The use of data
analysis on scientific publications facilitates and bring forward
information that could not been found with ancient data
analysis methods.

The research universe defined for this study is Malaria
disease contained in the so-called neglected tropical diseases
(NTD). The term NTD refers to a group of diseases that are
especially endemic in low-income populations living in
tropical and subtropical countries. However, up to this date,
there are no clear or agreed definitions of what a neglected
disease is constituted by. According to the 2010 report from the
World Health Organization (WHO) “Working to Overcome the
Global Impact of Neglected Tropical Diseases” the NTDs
affect more than 1 billion people around the world.

The known diseases are defined by the WHO as types I, II
and III where Malaria is considered a type III disease. The
definition includes the wealth of a country between rich and
poor, the state of its development between developed and
developing and most importantly a measure of the burden of
diseases by the incidence of the disease within the population
(CEWG 2012). The definitions themselves are combined such that:

- Type I diseases: are incident in both rich and poor
countries, with large numbers of vulnerable
populations in each, example cancer.
- Type II diseases: are incident in both rich and poor
countries, but with a substantial proportion of the cases
in poor countries, example tuberculosis.
- Type III diseases: are those that are overwhelmingly or
exclusively incident in developing countries, example
malaria.

With regard to investments in the area, those diseases do
not receive adequate attention for its treatment and eradication,
which perpetuates the condition of people who are exposed to
them [1].

It’s imperative that new tools and approaches being
developed focus on facilitating public policy planning and
foster the management of innovation in countries’ public health
systems [2]. Recent studies on Leishmaniasis [3] and Chagas
disease [4] used the Medline database and helped us with the
understanding and alignment of our work. A comparison with
results for co-authorship research [5] has also been made in
regards to the same database analyzing the relationships within
the scientific community.

According to the World Malaria Report [6] in 2010 there
were an estimated 219 million cases and 660,000 deaths.
Malaria is present in over 100 countries and threatens half of
the world’s population. In sub-Saharan Africa malaria is the
single largest cause of death for children under five. However
there is an expectation of 75% reduce in malaria cases by 2015
compared to year 2000 due to investments and collaborative
work. Investments in the area have grown significantly in the
past years from less than US$ 100 million in 2000 to US$ 1.71 billion in 2010. The increased investments have their consequences on the significant growth of research on the area. Figure 1 shows a world map of malaria endemic countries such that the countries in orange and red are the most affected ones.

II. DATA ANALYSIS OF A NEGLECTED DISEASE

The use of data analysis for neglected diseases through bibliometric or social network on scientific publications is a way to help managers on different positions cope with the necessity to view and take decisions on those areas based on reliable information regarding research. The use of tools and technologies on those areas has helped to find and manage information on large sets of data.

A. METHOD FOR DATA TREATMENT

For the data treatment and analyses we used the platform software Tetralogie. The software allows several and remote users to conduct strategic analysis from heterogeneous textual data through analysis of conventional and innovative methods. With statistical methods and exploratory data analysis methods, it shows strategic information such as actors identity, mobility, emergence and evolution of topics, concepts and terminology. Another feature of the software is that it allows graphical visualization for understanding human activities and their interactions as well as their evolution in a decision-making perspective. It combines static and dynamic visualization where the static aspect is based on a representation space, in which the precepts of graph theory are applied. The dynamic visualization applies features and capabilities to further understand evolution of the data and facts in respect to time. Tetralogie allowed us to identify the role of data within the structure, analyzing their neighborhood, filtering, k-core, transitivity, back to the source documents, to partition the graph or to focus on its structural features [7].

B. DATA RETRIEVED

For this study we have used the PubMed database, which accesses the Medline database. PubMed is a free database accessing references and abstracts on life sciences and biomedical topics. The access to the database occurred during the months of May and June, 2013. We have decided to take into consideration only the publications and patents from 2000 to 2012 included.

The research has been made taking into consideration authors of articles, countries where those authors were from, journals used for the scientific publications besides MeSH (Medical Subject Headings). Data files were retrieved from PubMed with the term(descriptor “Malaria” for title and abstract.

In regards to the number of publications for the years studied we have retrieved 27,880 articles containing a total of 51,720 authors from 121 different countries. The number of journals added up to 2,330, however only 412 of those had 10 or more articles in the area. The number of MeSH words filed on those articles were 8,890 from which the top 09 MeSH words were mentioned more then 10,000 times.

III. NETWORK ANALYSIS

The pattern of relationships of a system can be represented as a network where system components are the vertices and the connections between them ties. The structure of these networks and the specific pattern that might arise from its analyses has a major impact on the behavior of the system and its informational structures. Understand patterns in existing socio-technical networks and to intervene in these networks is one of the possibilities raised by this study.

The Network Analysis or Social Network Analysis (SNA) includes an extensive set of methods for assessing the structure of social groups [8], [9]. The use of SNA enables differentiated perception of a social environment. With this tools a new phenomena can be investigated and new hypotheses can be established. The SNA is derived from both the social sciences and the exact sciences, and their applications are enhanced by the ability to compute mathematical calculations and modeling as well as the expansion of communication networks and the use of digital environments.

The SNA brought interest from researchers in different fields of knowledge. In attempt to understand the impact of the relationship structure on networked environments, SNA gave rise to several analysis methods, which were based on the relationships between individuals or other type of elements, in a structured shaped network.

A social network consists of one or more finite sets of actors (vertices) and the defined relationships between them. The actors in a social network can be either individuals or companies, analyzed as individual units, as well as collective social units, for example, departments within an organization, public service agencies, nation-states of a continent or the world. The concept of actor is flexible, allowing different levels of aggregation, which allows its adaptation to different research problems [9].

A network is a simplified representation of a system that is reduced to an abstract structure consisting largely of patterns of relationships among the vertices. The vertices and lines can have extra information like names or size so that more details about the system can be captured. However, even with this extra information still much information is lost in the process of reducing an entire system on a network representation [5].

Besides these scholars, work done in the area of biotechnology and in the area of neglected diseases in NTDs endemic countries has been a reference to this scientific work. Dr. Carlos Morel Medici has published a series of articles using the methods of SNA to relate co-authorship in scientific research [1]. In France Dr. Brigitte Gay has conducted research on partnerships in the biotechnological industry and how these partnerships can define the success or failure of their companies [11], [12].

IV. RESULTS

The first step on our study was to understand the growth in regards to the numbers of publications and authors per year as seen in figure 2.

As can be seen on the figure, more publications have a direct relation to increased number of authors. With more and
new researchers involved the area tend to be more dynamic and new finds could help solve issues related to the disease. The total number of publications and authors has increased steadily the past 13 years which might collaborate with the investments in the area and the reduce expectation of malaria cases. The total number of publications has almost double on those years in the area and the reduce expectation of malaria cases. The new finds could help solve issues related to the disease. The new researchers involved the area tend to be more dynamic and new finds could help solve issues related to the disease.

### A. COUNTRIES, JOURNALS AND AUTHORS PUBLICATIONS

In regards to country publications we can see a growth in interest, which has more than doubled the number of publications since 2000 in most countries but more so on countries with endemic problems such as India, Thailand, Nigeria, Kenya and Brazil. Those countries are becoming more involved in research, and due to its close relation to the disease there is an urge to find solutions. However, countries such as USA, UK, France and Australia still play an important role on research and publications. The financial support received by those countries on research and technologies has helped millions of people. Figure 3 shows an overview of the world map in terms of publications on PubMed. The countries with a lighter green have more publications than countries with darker green. In the past few years countries such as India, Brazil and China, which are part of the “BRICS”, have had an important role over those researches.

As for the authors we can notice on table 2 that some of them have a higher number of publications, much greater than the average researcher on the area. The top 10 authors for example have more than 150 publications (11 are shown in the table) over the period while the number of publications per author is 2.3. This demonstrates a high concentration or centrality in a network. One problem that we might encounter in researching over authors publications is theirs alias or different formats for ones own names such as smith t or smith t . a which might characterize the same person. To partially resolve the problem, Tetralogie creates a synonymous table for authors and other elements as needed. For the data analysed there were a table with 7,036 author synonymous, so for those authors all their publications were added up to one name.

The understanding of the participation of those authors during the period is also important. Figure 4 shows an analysis on individual authors called “Histogram of the Individual” over the period from 2000 to 2012, where the red bar demonstrate the number of publications for that particular year. As an example we have chosen authors with different characteristics such as author that has just come into the field in regards to the area and the number of publications for that particular year. As an example we have chosen authors with different characteristics such as author that has just come into the field in regards to the area and the number of publications for that particular year.

### Table 1 – Top Journals by the number of publications per year

<table>
<thead>
<tr>
<th>Journals</th>
<th>Number of Publications per Year (2012 – 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malaria Journal</td>
<td>518</td>
</tr>
<tr>
<td>American Journal of Tropical Medicine</td>
<td>429</td>
</tr>
<tr>
<td>Plos One</td>
<td>387</td>
</tr>
<tr>
<td>Trans of the Royal Society of Tropical Medicine and Hygiene</td>
<td>350</td>
</tr>
<tr>
<td>Journal of Infectious Diseases</td>
<td>343</td>
</tr>
<tr>
<td>Infection and Immunity</td>
<td>343</td>
</tr>
<tr>
<td>Acta Tropical</td>
<td>322</td>
</tr>
<tr>
<td>Trends in Parasitology</td>
<td>310</td>
</tr>
<tr>
<td>Lancet</td>
<td>308</td>
</tr>
</tbody>
</table>

### Table 2 – Author Publication

<table>
<thead>
<tr>
<th>Journals</th>
<th>Number of Publications per Year (2012 – 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White, N</td>
<td>310</td>
</tr>
<tr>
<td>Looareesuw</td>
<td>231</td>
</tr>
<tr>
<td>Kremersn, P</td>
<td>208</td>
</tr>
<tr>
<td>Marsh, K</td>
<td>193</td>
</tr>
<tr>
<td>Snor, R</td>
<td>192</td>
</tr>
<tr>
<td>Doumbo, O</td>
<td>182</td>
</tr>
<tr>
<td>Nosten, F</td>
<td>175</td>
</tr>
<tr>
<td>Rosenthal, P</td>
<td>168</td>
</tr>
<tr>
<td>Greenwood</td>
<td>159</td>
</tr>
<tr>
<td>Alomso, P</td>
<td>151</td>
</tr>
<tr>
<td>Smith, T</td>
<td>138</td>
</tr>
</tbody>
</table>

Figure 2 – Authors and Publications per year

Figure 3 – Country Publication
and 2008); Or authors which have been consistent and present during the entire period (published all years from 2000 to 2012).

This type of analysis can be used to verify individual behaviours on specific authors, such patterns on publications.

B. SOCIAL AND MORPHOLOGICAL NETWORKS ANALYSIS

As we have mentioned on this study new forms of analysis must take place in order to better understand the information that is available. One of those forms is the network analysis, which consists of an analysis over the relationship amongst people or words over those publications. The most common and already in use over scientific publications is the social network analysis where authors, which have co-author a paper, demonstrate a relationship and so can be analyzed as such. One important issue related to visual analysis is the ability to understand what has been presented, which might become a problem when we are dealing with large sets of data such as the case of scientific publications. Before we can make any sense of what might appear on a network graph it is necessary to sort the data that is most important or relevant at the moment.

Figure 5 shows 3 tables for an author relationship with a lower density for the ones in the right compared to the previous one in the left. The left most table on figure 5 shows the authors with 5 or more publications, over the period study, and its relationships. An author defined on the x axis, that have co-author a paper with another author, on y axis, will have a red dot over its column and on the respective line for its companion. Once we were dealing with more than 50,000 authors, the size of the table, without the correct filters might become an issue. Noticed that on that figure there is a great density where most of the authors have relationships with several of its peers. The red line crossing the center is where the author relates to itself on the x and y axis and is not considered. Because the table is a symmetric matrix all the dots on the left of the center line will appear on the right side of that line.

One of the analyses for that graph is the clustering coefficient, which were 0,496 for the Malaria co-authorship. The clustering coefficient measures network “clustering” or “transitivity,” which is the probability that two of a scientist’s coauthors have themselves coauthored a paper. In science a high clustering coefficient would suggest that people are working together quite often, while a very detached network, as in patents for example, might suggest that people very rarely contribute with people outside its group. This demonstrates that on the malaria research environment people tend to collaborate quite often with each other. This is also true for other areas of research [5].

The middle table shows the relations amongst authors that have worked together on at least two occasions (publications) and the right most table authors that have worked together on at least three occasions. From the rightmost table we focus our analysis on the area where there is a greater concentration (top left) of relationships and from there we draw figure 6.

The social network analysis consists of understanding the connections (relations) existent over the network. Each red dot on figure 6 denotes a researcher and each line a relationship with another researcher. People with more connections have worked with more people and in consequence they have a higher centrality. Some interesting connections happen also amongst clusters. On Figure 6 we have 3 clusters that are connected with each other by few people. Those researchers might not have a higher centrality degree (absolute number of connections) but are key to keep the network together and help the flux of information. For this network we are showing authors that have at least 5 publications over the period and that have worked together on publications on three different occasions. The graph is only a partial view of the third table on figure 5.

A number of conclusions could be made with a simple visualization analysis over the graph. The authors, for example, with higher centrality might have the ability to influence a greater number of people. On the other hand, those same authors are exposed to a great number of ideas. As for the researchers that make the connections between clusters we can say that they might work as gate keepers and are the ones responsible to share new ideas to their own group or close related friends. A closer study over each individual author and its own work is necessary to further understand the possibilities and implications of such a graph but for the purpose of this work we have only raised the issue without a deeper analysis. The graph helps us to pinpoint the most important players and the ones we should pursue a further study.
used data and network analysis tools on scientific databases. This line of research have been made by different authors before and now we aimed to put together some of the techniques used.

As for the findings on the field we can say that there have been an increase interest by the scientific community over the years by the increase number of publications and a steady growth of new scientists on the area. In regard to the countries investing on the area, new players are coming forward, especially among the countries with high rate of neglected disease incidence, such as India. As for the network analysis we could see that there is a high collaboration due to its clustering coefficient, which is in accordance with the scientific communities. Within those authors we found some key players in terms of the number of publications and those players might be a point for a further study. This study however has not shown deeper studies over specific authors evolutions using network analysis or morphological evolution using a similar methods. Those studies are part of our current work and might be demonstrated on future analysis.

**BIBLIOGRAPHY**


---

**V. CONCLUSION**

This study was an attempt to apply some methods to analyze data focusing on Malaria, a neglected disease. We have...