Social network analysis and mining: challenges and applications

Bruna P. Fonseca[1], **Priscila C. Albuquerque**[1], **Fabio Zicker**[1], **Carlos M. Morel**[1]

[1]Centro de Desenvolvimento Tecnológico em Saúde (CDTS)
Fundação Oswaldo Cruz (Fiocruz) Rio de Janeiro – RJ – Brazil

(bruna.fonseca, priscila.costa, fabio.zicker, carlos.morel)@cdts.fiocruz.br

Abstract. *Social network analysis and mining (SNAM) is a powerful tool to disclose relevant information hidden in large volumes of raw data. Its application to several research fields, powered by automation and advanced computing infrastructure, expanded its use and brought along new challenges. In this paper, we provide a critical perspective on SNAM's major challenges, by discussing a few examples. We also address some promising applications that can potentially translate SNAM results into practical knowledge.*

Resumo. *A análise de redes sociais e mineração de dados (SNAM) é uma ferramenta poderosa que revela informações relevantes que estão ocultas em grandes volumes de dados brutos. Sua aplicação em diversos campos de pesquisa, impulsionada pela automação e infraestrutura computacional avançada, expandiu seu uso e trouxe novos desafios. Neste artigo, fornecemos uma perspectiva crítica sobre os principais desafios da SNAM, por meio da discussão de alguns exemplos. Também abordamos algumas aplicações promissoras que podem potencialmente traduzir os resultados da SNAM em conhecimento prático.*

1. Introduction

Social network analysis and mining (SNAM) is a powerful tool that can disclose relationships, patterns, and other relevant information hidden in large volumes of raw data. For the past decade, SNAM research has advanced and expanded, as seen by the increasing number of publications using this method (Figure 1) and diversity of research fields in which it has been applied (Figure 2). SNAM developed into a comprehensive multidisciplinary field, attracting the attention of scientists, scholars, government agencies and industries. Scientific publications applying a network perspective now appear in major multidisciplinary journals, such as *Science* and *Nature*, and in at least six peer-reviewed journals exclusively dedicated to network analysis: *Social Networks, Social Network Analysis and Mining, Journal of Social Structure, Network Science, Connections*, and the *Journal of Complex Networks*.

SNAM has a solid foundation in science and, because of the significant developments in information technology, it has become popular in the traditional scientific community. From data collection to network analysis, automation and advanced computing infrastructure enabled researchers to process and analyze large and diverse data, thereby extending network analysis to nearly hundreds of thousands of data points. Recent improvements in the methods of describing, summarizing, comparing and quantitatively

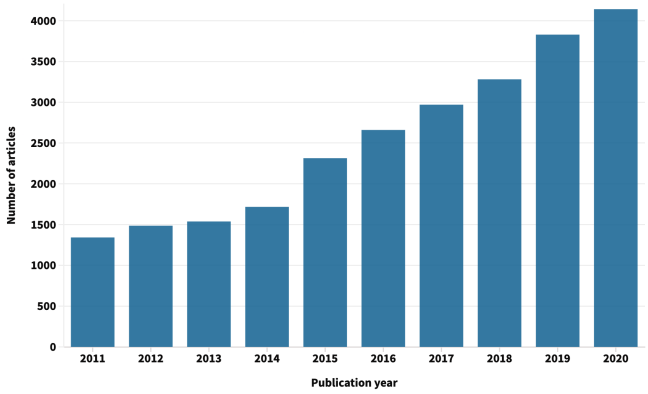

Figure 1. Number of scientific publications using SNAM (2011-2020). Publications were retrieved from the Web of Science database, searching for "social network analysis" or "data mining" or "text mining" in the titles or abstracts of publications. Only articles were considered in the analysis (n= 25,903).

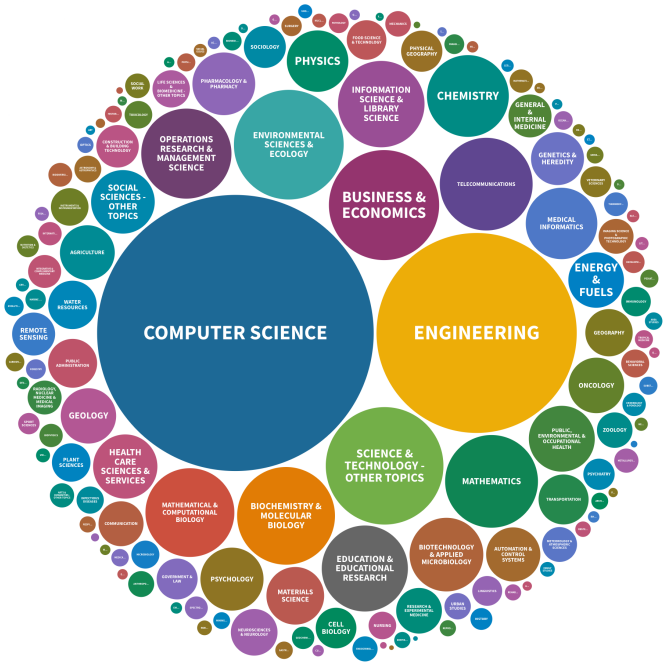

Figure 2. Research fields using SNAM (2011-2020). Publications were retrieved from the Web of Science (WoS) database, searching for "social network analysis" or "data mining" or "text mining" in the titles or abstracts of publications. Only articles were considered in the analysis (n= 25,903). Research fields were identified based on WoS categories. The size of the circles is proportional to the number of articles in each research field.

analyzing social structures, relationships and networks have also influenced the use of SNAM.

The Brazilian Workshop on Social Network Analysis and Mining (BraSNAM) has provided an interdisciplinary venue that brings together scholars and researchers in related fields to promote collaboration and exchanges. To celebrate the 10th anniversary of BraSNAM, we will briefly introduce some major challenges and potential applications of SNAM research.

2. Challenges

2.1. Absence of readily available sources of structured data

SNAM's demand for high-quality data is indisputable. Due to the almost unlimited sources of data, types of analysis and applications, the availability and structure of data have become more and more important. Incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data can greatly affect the results.

The analysis of research funding is a good example of the indispensable need of accessible structured data. Budget constraints and changes in political priorities have long challenged the sustainability of research funding. The analysis of the relationship between funding and research outputs has important implications for understanding the efficiency of funding mechanisms and the relevance of research. Brazil has multiple funding agencies (state, national, private, philanthropic), and a large portfolio of projects, but there is no structured, interoperable, publicly available, common source of information. Researchers could explore the national funding landscape data to make sense of collaborations, leadership, overlapping and complementary interests, generating valuable insights to inform the national research funding policies. Unfortunately, the information (when available) is dispersed in several unrelated databases, with different formats and frequently missing data. Building nationally oriented, readily available and interoperable databases is key to leverage all the potential of SNAM to uncover strategic information.

2.2. Time-consuming data collection and cleaning

In the early days of SNAM, data collection was basically a small-scale, direct hands-on process, from homogeneous and small datasets. With the vast amount of data currently available for analysis, manual data cleaning became almost impossible. Automated collection of raw data has made things easier, but the diversity and heterogeneity of data brought new challenges.

Data cleaning involves eliminating the errors, resolving inconsistencies and transforming the data into a uniform format. An example of the importance of data cleaning is the analysis of co-authorship networks. To ensure a more accurate and complete picture of the research collaboration landscape, co-authorship network analysis often requires collecting data from different international or regional databases. The inclusion of preprint repositories has also become of great value in addressing cutting-edge research. Integrating these different databases by pairing matching fields, removing duplicate records, eliminating ambiguities of institution and/or author names, and converting final data into relational data suitable for network analysis can be a time-consuming task. Innovations that remove or mitigate this issue are research opportunities that have the potential to simplify the steps that precede and follow network analysis.

2.3. Non-critical use of network visualization and descriptive metrics

The popularity of visualization software has fueled the non-critical use of network visualization and descriptive metrics, which may be detached from the real world. Different network structures are suitable for different purposes. Therefore, due to the lack of an "ideal" network, it is impossible to establish a standard benchmark for evaluating networks. Additionally, network metrics based on snapshots are of limited value. Since networks are often highly dynamic, data on multiple time points is required to understand their structure, evolution and progress.

Density is an example of a largely misused metric in SNAM research. The network density is very sensitive to changes in network size, thus not suitable for comparing networks of different sizes. In collaboration networks, as the network expands over time, its density tends to decrease due to inherent difficulties of maintaining effective collaboration. Collaboration relies heavily on people making personal connections. Low density in large-scale collaboration networks is a common feature and does not necessarily mean a decrease in cooperation.

2.4. Need of a theoretical basis to contextualize and interpret results

Although SNAM has become increasingly popular, the interpretation of results is not always that simple. The methods themselves do not require or imply any particular theory, but they do require theoretical contextualization in broader debates. Unfortunately, it is not yet a common practice to justify the technical choice of network metrics or explain their meaning in a specific research setting.

The solid interpretation of research results requires both theory and context. This kind of opportunity should be used for interdisciplinary cooperation. Researchers can cooperate productively with SNAM experts to conduct robust analysis and explanatory studies that can further promote the SNAM agenda.

2.5. Lack of specialized human resources

Although SNAM is an expanding field of research, there is still a lack of professional human resources capable of dealing with data science, visualization and big data, specially in low- and middle-income countries (LMIC).

Strengthening individual and institutional SNAM capacities in LMICs would provide an enormous opportunity to support large-scale activities, from monitoring migration flows [Windzio et al. 2021] to powering criminal intelligence activities [Cunha and Gonçalves 2018] and supporting complex project management [Lee et al. 2018]. In addition, LMIC-based scientists could further participate in international research bringing local views and research questions that reflect their needs. The creation of a center of excellence in SNAM, comprised of highly skilled individuals with leading-edge knowledge and competency in this field, could offer support and training in this area.

2.6. Applications

Although SNAM provides insightful perspectives into the relationships and structural characteristics of networks in many research fields, there are relatively few discussions on how to translate results into applied knowledge. Here are a few examples of promising areas that SNAM has contributed in generating essential information

2.7. Evaluation of research programs

SNAM is a useful tool for evaluating research programs in which relationships are important to produce effective results. In emerging research fields in which scientific knowledge is still maturing, the traditional indicators (such as number of scientific articles and impact factor of the journals) are of limited value. New indicators and standards based on SNAM could allow for a fair and effective evaluation without ignoring scientific standards.

This issue was addressed in a national research program on neglected tropical diseases in Brazil [Morel et al. 2009]. Publications related to these diseases published by Brazilian researchers were retrieved from international databases, analyzed and processed with SNAM tools to build a co-authorship network. Network component analysis generated a picture of the overall network structure, identifying fragmented areas and possible weaknesses. The analysis revealed institutions acting as network cut-points that were critical key players, being responsible for maintaining other institutions from less-developed regions connected to the network. These institutions were considered fundamental partners for training, capacity building and research strengthening. The innovative contribution brought by SNAM allowed the assessment of the evolution, performance and robustness of the networks involved in the program, showing opportunities for strategic management.

2.8. Institutional management

Network analysis can explore and produce useful information about collaborations and partnerships, supporting the performance assessment and development of research institutions [Fonseca et al. 2016]. By mapping the pattern of information flow across research groups within institutions, SNAM can yield critical insights to promote collaboration that will provide strategic benefit for these institutions. SNAM results can be considered an integrative reference for establishing action plans and supporting management decisions.

SNAM was used to examine the evolution and dynamics of the Brazilian research network on tuberculosis, a strategic research area for the Oswaldo Cruz Foundation (Fiocruz) [Fonseca et al. 2017]. The analysis was helpful in identifying Fiocruz's most influential researchers, which could serve as advisors/experts for investment and induction policies, as well as leading researchers that could improve information exchange, systems integration and innovation in the institution. The results suggested a limited interaction within the tuberculosis research field at Fiocruz and provided inputs for planning actions to promote synergy between internal research groups working in complementary areas.

2.9. Prevention and control of epidemics

SNAM can also provide an important tool to analyze the role of countries and their institutions during sanitary crises. During epidemics, the position occupied by countries, institutions and authors in a knowledge-generating network is an important parameter for influencing response, decision-making, preparedness and empowerment. The analysis of the respective research networks established during the Ebola epidemics in West Africa (2015) and the spread of Zika virus in South America (2016) revealed different profiles in terms of the geographic location of the authors and the relevance of the affected countries in co-authorship networks [Vasconcellos et al. 2018]. While Brazil was among the most central countries in the Zika research network, with national researchers being responsible

for seminal work on outbreak characterization and clinical case definitions, the majority of the work conducted in West Africa to detect, diagnose and control the Ebola epidemics was carried out by international teams from abroad.

The ongoing Covid-19 pandemics also brought new opportunities for SNAM research in epidemiology and global health issues. The detection, mapping and tracking in real time superspreading events and the emergence of SARS-CoV-2 Variants of Interest (VOI), Variants of Concern (VOC) and Variants of High Consequence (VOHC), provide critical information to cope with the spread of the disease [CDC 2020].

2.10. Policy planning and innovation management

Analysis of scientific publications and patents is a quantitative method widely used to examine the knowledge structure and technological development of research fields. Comprehensive mapping of a specific research area through SNAM can provide useful insights and facilitate the incorporation of data into the planning and policy-making processes, promoting innovation management.

SNAM was used to generate evidence on microscopy-related research in Brazil [Albuquerque et al. 2019], a key area in the national scientific and technological scenario. The authors have shown that the research community has been growing over the past 10 years, and pointed out that knowledge was concentrated in the southeast region of the country, based on an intra-regional collaboration pattern. The study acknowledged the need for the expansion and promotion of nation-wide collaboration networks. Innovative strategies and policies can foster the engagement of research organizations outside the "southeast hub" to also benefit from the intellectual and financial leverage of these partnerships. The most central institutions identified could facilitate the development and implementation of a national network in order to harmonize and strengthen national research capacity.

SNAM was applied as a new tool to facilitate public policy planning and innovation management in tuberculosis research in Brazil [Vasconcellos and Morel 2012]. Through the joint analysis of scientific publications and patents, the authors have identified most prominent researchers, responsible for consolidating the scientific knowledge on the disease as well as leading technological innovation. The results endorsed the importance of maintaining the continuity of national production development policies and infrastructure support to transform the potential of research into public health benefits.

3. Conclusions

The future of research and practice in the field of SNAM is both promising and challenging. Opportunities are many: theoretical and applied research has been published in specific journals as well as traditional venues. Based on our experience at the Center for Technological Development in Health (CDTS) at Fiocruz, we highlighted important challenges related to data availability, access, collection and cleaning; need of solid theoretical basis and contextualized use of metrics; and shortage of specialized personnel. We believe that the fields of evaluation, management, control of epidemics, policy planning and innovation management are areas in which SNAM can provide useful information for action. We encourage students and researchers to further explore the full range of opportunities offered by SNAM, bearing in mind that interdisciplinary cooperation could significatively enhance the research agenda.

References

Albuquerque, P. C., de Paula Fonseca E Fonseca, B., Girard-Dias, W., Zicker, F., de Souza, W., and Miranda, K. (2019). Mapping the Brazilian microscopy landscape: A bibliometric and network analysis. *Micron (Oxford, England: 1993)*, 116:84–92.

CDC (2020). Cases, Data, and Surveillance.

Cunha, B. R. d. and Gonçalves, S. (2018). Topology, robustness, and structural controllability of the Brazilian Federal Police criminal intelligence network. *Applied Network Science*, 3(1):1–20. Number: 1 Publisher: SpringerOpen.

Fonseca, B. d. P. F., Fernandes, E., and Fonseca, M. V. A. (2016). Collaboration in science and technology organizations of the public sector: A network perspective. *Science and Public Policy*, 44(1):37–49.

Fonseca, B. d. P. F. E., Silva, M. V. P. d., Araújo, K. M. d., Sampaio, R. B., and Moraes, M. O. (2017). Network analysis for science and technology management: Evidence from tuberculosis research in Fiocruz, Brazil. *PloS One*, 12(8):e0181870.

Lee, C.-Y., Chong, H. Y., Liao, P., and Wang, X. (2018). Critical Review of Social Network Analysis Applications in Complex Project Management. *Journal of Management in Engineering*, 34(2). Accepted: 2018-02-01T05:24:34Z Publisher: ASCE.

Morel, C. M., Serruya, S. J., Penna, G. O., and Guimarães, R. (2009). Co-authorship network analysis: A powerful tool for strategic planning of research, development and capacity building programs on neglected diseases. *PLoS Neglected Tropical Diseases*, 3(8):e501.

Vasconcellos, A. G., Fonseca, B. d. P. F. e., and Morel, C. M. (2018). Revisiting the concept of Innovative Developing Countries (IDCs) for its relevance to health innovation and neglected tropical diseases and for the prevention and control of epidemics. *PLOS Neglected Tropical Diseases*, 12(7):e0006469.

Vasconcellos, A. G. and Morel, C. M. (2012). Enabling policy planning and innovation management through patent information and co-authorship network analyses: A study of tuberculosis in Brazil. *PLoS ONE*, 7(10):e45569.

Windzio, M., Teney, C., and Lenkewitz, S. (2021). A network analysis of intra-EU migration flows: how regulatory policies, economic inequalities and the network-topology shape the intra-EU migration space. *Journal of Ethnic and Migration Studies*, 47(5):951–969. Publisher: Routledge _eprint: https://doi.org/10.1080/1369183X.2019.1643229.

Corresponding Authors: Photos and mini-biographies

Bruna Fonseca

Biologist at the Federal University of Rio de Janeiro (UFRJ), with specialization in social network analysis from the University of Greenwich (UK) and a PhD in Production Engineering at COPPE/UFRJ with emphasis in Knowledge Management. Served as a member of the Management Committee of the Fiocruz Observatory in Science, Technology and Innovation in Health and as a member of the Technical Groups of Technological Development and Innovation Indicators and Collaboration Networks. Currently coordinates the area of Networks in ST&I in Health at the Center for Technological Development in Health (CDTS/Fiocruz) with a research focus on the application of network analysis as a support tool for the planning and management of STI, including the mapping of topics of interest in public health, studying the dynamics of scientific and technological collaboration and the relationship between scientific production and societal needs.

Carlos Medicis Morel

CNPq Emeritus Researcher, Full Member of the Brazilian Academy of Sciences and The World Academy of Sciences (TWAS). Medical Doctor at the Federal University of Pernambuco (UFPE) and Doctor in Science at the Carlos Chagas Filho Biophysics Institute of the Federal University of Rio de Janeiro (UFRJ). Former President of the Oswaldo Cruz Foundation (Fiocruz) and former Director of the Special Program for Research and Training in Tropical Diseases (TDR) of the World Health Organization (WHO) in Geneva. Currently Coordinator of the Center for Technological Development in Health (CDTS) at Fiocruz; Member of the Board of Directors of FIND, Foundation for Innovative New Diagnostics, in Geneva; and Consultative Member of the Harvard University/WHO Advisory Committee *Rethinking Malaria in the Context of COVID-19*.