A Network-based Approach to Evolution of MEDLINE

Andrej Kastrin*, Thomas C. Rindflesch† and Dimitar Hristovski‡
*Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia
Email: andrej.kastrin@mf.uni-lj.si
†Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, USA
Email: trindflesch@mail.nih.gov
‡Institute of Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia
Email: dimitar.hristovski@mf.uni-lj.si

***Abstract*—MEDLINE bibliographic database can be represented as a network of nodes and edges, where the former represent biomedical concepts and the latter represent relationships among them. Nodes and edges are not uniformly distributed but rather appear in locally dense communities. We investigate the dynamics and evolution of MEDLINE using network analysis based on community modeling. This study identifies the major research focuses and the current status and trends in the life sciences. To the best of our knowledge, this is the first analysis conducted on such a large portion of the MEDLINE database.**

Keywords–Complex networks; Network analysis; Network evolution; MEDLINE.

I. INTRODUCTION

The growth of science is increasingly dynamic and interdisciplinary. Barriers (silos), both physical and conceptual, that once effectively isolated researchers are breaking down. One consequence of this is that the biomedical literature is large and complex, and the number of published papers is growing at a considerable rate. It is therefore becoming ever more important to identify and describe developing research trends and to follow their evolution over time.

The premier repository of research in the life sciences is the MEDLINE bibliographic database. It contains over 24 million citations, with around 4000 citations added daily. MEDLINE can be represented as a network of nodes and edges, where the former represent biomedical concepts and the latter represent relationships among them, as we have shown in previous research [1]. Such a network can represent the structure and dynamics of a complex system; topological properties of the network can both elucidate static patterns and predict the future by computing changes over time. Co-occurrence in the network between concepts such as genes, diseases, biological processes, or chemical compounds represents biomedical knowledge.

When a network represents the real world, its nodes and edges are not uniformly distributed but rather appear in locally dense clusters, or communities, embedded within the larger structure [2]. Community structures are often of particular interest as forming the basis of network analysis aimed at elucidating knowledge represented by the network. A community is defined as a subset of nodes sharing similar properties and recognizable as being distinct from the larger network. A complex network representing the real world (such as that representing MEDLINE) is an evolving structure that change over time either by adding new nodes or by forming new relations between existing nodes [3]. This expansion can proceed over time with considerable speed in terms of size and space over time.

Science as a complex system has been studied from the point of view of network analysis, for example through co-authorship network analysis [4] or single-word analysis [5]. In a related approach, network analysis of the development of scientific knowledge may be valuable in building theoretical models of the collective dynamics of science. Here, we investigate the dynamics and evolution of MEDLINE using network analysis based on topic modeling. Our hypothesis is that the history of the biomedical domain can be summarized in a series of co-occurrences of keywords (i.e., Medical Subject Headings (MeSH) terms) that are associated with each MEDLINE citation and that identify its important topics. MeSH is a controlled vocabulary made up of biomedical terms at several levels of specificity. We model scientific topics as evolving communities of MeSH terms over time and explore how the temporal characteristics of the MeSH network can be used to provide insight into the historical evolution of scientific thought in biomedicine. As a proof of concept, we implemented a computational experiment based on over 20 million documents in MEDLINE, from 1966 through the end of 2014. To the best of our knowledge, this is the first analysis conducted on such a large portion of the MEDLINE database. In particular, this analysis is of special interest to researchers who seek to acquaint themselves about scientific topics, trends, and collaboration opportunities.

The abstract is structured as follows. Section II describes methodology of this study and Section III provides some empirical evidence. Conclusions and the scope for future research are presented in Section IV.

II. METHODS

We processed MEDLINE from 1966 up to the end of 2014, only including citations tagged with major MeSH descriptors. In the constructed network, nodes represent major MeSH descriptors, and edges between two descriptors represent co-occurrence of those descriptors in the same MEDLINE citation. Recognizing the critical events that describe changes in network structure over time constitutes one way of tracking the development of communities. To capture critical events, we converted the entire network into static subnetworks at yearly snapshots, from 1966 through 2014. These subnetworks embed the communities (groups of nodes) of MeSH terms through which their development over time can be observed. We used the Louvain community detection algorithm to identify communities of nodes in each subnetwork [6]. After discovering communities, we computed relationships between them, in order to track their evolution over time. The Jaccard coefficient was first used to determine whether two

 ISBN: 978-1-61208-558-6

communities match [7]. Next, we characterized the content of each community by computing its density and centrality. Density measures both the strength of the edges that tie the cluster of MeSH terms together, as well as the clusters capacity to develop ever time. Centrality measures the degree of interaction of a cluster with other parts of the network. Finally, we created a strategic diagram for each time slot, which is a graphical representation of the structure of the particular scientific field. We can identify four types of clusters according to the quadrant in which they appear in the diagram: (i) motor themes, which are both well developed and important for a research field; (ii) specialized and peripheral themes; (iii) themes, which are either emerging or disappearing; and (iv) themes, which are important for a research field but are not developed.

III. Results

This study identifies the major research focuses and the current status and trends in the life sciences. Overall, we extracted about 1700 different evolving communities. Figure 1 depicts heat map of evolving communities for various timeslots. As the shade of cell darkens, the size of the community increases. We could observe interesting pattern from the plot; it seems that each community build its existence on the basis of previous communities, therefore the trend line is diagonal with rare exceptions that extends over large periods of years.

Next, in this study we provide a description of the intellectual structure and dynamics of the entire field of biomedicine from the perspective of frequently appearing MeSH descriptors. The results of this study show that (i) using MeSH terms is plausible for tracking historical events in the biomedical domain; (ii) the evolution of MEDLINE occurs in an incremental fashion; (iii) over the years increasingly diverse research disciplines are involved in the complex process of scientific evolution, and links among them become stronger; and (iv) different research areas have different dynamic evolution patterns.

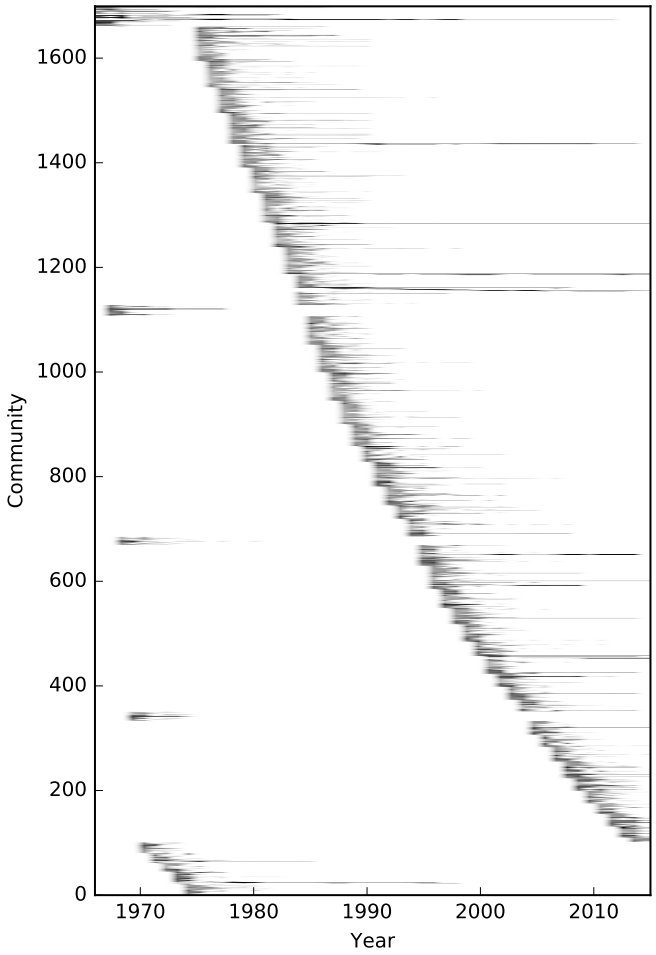

Figure 1. Heatmap of evolving communities for each timeslot

IV. Conclusion

When compared to existing research, this work is innovative in three respects: (i) the experimental design incorporates the longitudinal framework based on dynamic communities, (ii) we provide visualization of the evolution of research topics, and (iii) we propose a simple community labeling approach based on MeSH terms. There are also many opportunities for future work. First, we should address the problem of filtering co-occurrences. Second, we should address automatic cluster labeling, while manual labeling is tedious, time-consuming, and expensive task. Our long-term interest is also to include temporal frequent pattern mining into the analysis.

References

[1] A. Kastrin, T. C. Rindflesch, and D. Hristovski, "Large-scale structure of a network of co-occurring MeSH terms: Statistical analysis of macroscopic properties," PloS One, vol. 9, 2014, p. e102188.

[2] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," Phys Rev E, vol. 68, 2003, p. 36122.

[3] P. Holme and J. Saramki, "Temporal networks," Phys Rep, vol. 519, 2012, pp. 97–125.

[4] M. E. J. Newman, "The structure of scientific collaboration networks," Proc Natl Acad Sci USA, vol. 98, 2001, pp. 404–409.

[5] T. Kuhn, M. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," Phys Rev X, vol. 4, 2014, p. 41036.

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," J Stat Mech Theory Exp, vol. 2008, 2008, p. P10008.

[7] P. Jaccard, "The distribution of the flora in the alpin zone," New Phytol, vol. 11, 2012, pp. 37–50.

 ISBN: 978-1-61208-558-6