Journal of Informetrics 12 (2018) 757–783

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

Stochastic block model reveals maps of citation patterns and their evolution in time

Darko Hric, Kimmo Kaski, Mikko Kivelä*

Department of Computer Science, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland

ARTICLE INFO

Article history:
Received 2 May 2017
Received in revised form 30 May 2018
Accepted 30 May 2018

Keywords:
Web of science
Citation networks
Evolution of science
Stochastic block model

ABSTRACT

In this study we map out the large-scale structure of citation networks of science journals and follow their evolution in time by using stochastic block models (SBMs). The SBM fitting procedures are principled methods that can be used to find hierarchical grouping of journals that show similar incoming and outgoing citations patterns. These methods work directly on the citation network without the need to construct auxiliary networks based on similarity of nodes. We fit the SBMs to the networks of journals we have constructed from the data set of around 630 million citations and find a variety of different types of groups, such as communities, bridges, sources, and sinks. In addition we use a recent generalization of SBMs to determine how much a manually curated classification of journals into subfields of science is related to the group structure of the journal network and how this relationship changes in time. The SBM method tries to find a network of blocks that is the best high-level representation of the network of journals, and we illustrate how these block networks (at various levels of resolution) can be used as maps of science.

1. Introduction

The process of creating scientific knowledge relies on publications that are often stored and archived, with the primary purpose of preserving and distributing the knowledge obtained through research. These archives can also be used to study the science making itself, for example, by extracting information of collaborations, citations, or keywords of the published articles. Research in this field has a fairly long and rich history with wide range of research topics, like the assessment and prediction of performance and quality of individual papers, researchers, institutions, journals, fields, and even countries (Althouse, West, Bergstrom, & Bergstrom, 2009; Lehmann, Jackson, & Lautrup, 2008; Nerur, Sikora, Mangalaraj, & Balijepally, 2005), as well as identification of various large scale structures of science (Boyack & Klavans, 2014; Carpenter & Narin, 1973; de Solla Price, 1965; Leydesdorff, Carley, & Rafols, 2013; Small, 1999; Waltman, van Eck, & Noyons, 2010), journal classification (Janssens, Zhang, Moor, & Glänzel, 2009; Leydesdorff, 2006; Wang & Waltman, 2016; Zhang, Liu, Janssens, Liang, & Glänzel, 2010), following research trends (Chen, 2013; Persson, 2010; Porter & Rafols, 2009), and recognizing the emerging fields or researchers (Cozzens et al., 2010; Lambiotte & Panzarasa, 2009; Shibata, Kajikawa, Takeda, Sakata, & Matsushima, 2011; Small, Boyack, & Klavans, 2014; Small & Greenlee, 1989).

Bibliographic databases, like Web of Science, Scopus, and Google Scholar, store metadata of scientific publications, which can be used to analyse science making at all levels, from large scale structure to performance of individual papers. The number

* Corresponding author.
E-mail addresses: darko.hric@aalto.fi (D. Hric), kimmo.kaski@aalto.fi (K. Kaski), mikko.kivela@aalto.fi (M. Kivelä).

https://doi.org/10.1016/j.joi.2018.05.004
1751-1577/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

of *entities* in the data, including articles, journals, citations, and scientists is very large and keeps growing exponentially (Appendix A; Pan, Petersen, Pammolli, & Fortunato, 2016). To make sense of such massive amounts of data available about science one needs to simplify it and find its inherent patterns. This idea is not different from creating *maps* that provide a simplified description of reality, i.e. maps of science that describe the endeavour of science in a broad sense (Boyack, Klavans, & Börner, 2005; Chen, 2013; Small, 1999). Such a map needs to provide a reasonably accurate simplification of the structures it is mapping, i.e. individual elements need to be grouped (or clustered) to preserve large-scale patterns, while obscuring small and unimportant details. However, this is not a trivial task, and finding an optimal simplification accurately and reliably is becoming even more challenging as the networks under study continue to grow.

Conventional data analysis tools, such as clustering or dimension reduction methods, can be used to simplify the data about the complex relationships between the data entities. Representing the entities as vectors of their features is a common and practical abstraction that allows the use of clustering methods in the space of features, in which the most similar entities are grouped based on the similarity of the used features. These vectors can contain citation information between the entities, and one can define similarity measures, like bibliographic coupling, co-citation, distance between citation vectors (Euclidean, cosine, Jaccard, etc.), and correlation coefficients between the citation vectors or publication texts (abstracts, keywords, etc.) (Boyack et al., 2005; Carpenter & Narin, 1973; Janssens et al., 2009; Kessler, 1963; Leydesdorff & Rafols, 2012; Marshakova, 1973; Small, 1973; Wang & Koopman, 2017).

The data of scientific progress can be analysed with a variety of methods once the data has been preprocessed. The dimensionality reduction techniques project the vectors into the most significant subspaces revealing groups of correlated entities (multidimensional scaling, factor analysis, etc.) (Leydesdorff et al., 2013; Small, 1999). Classical clustering techniques, e.g. hierarchical clustering and k-means, operate on the full space of features, and provide clusters of similar entities, based on implicitly or explicitly defined similarity measure or distance (Boyack et al., 2005; Modha & Spangler, 2000; Punj & Stewart, 1983; Silva, Rodrigues, Oliveira, da, & Costa, 2013; Wang & Koopman, 2017). The factor analysis applied separately to the citing and cited direction of the complete citation matrix, enables further specialization into the types of groups it finds, since by using only one direction at a time, it detects groups based on past and future citations, separately (Leydesdorff & Rafols, 2009). The co-citation and bibliographic coupling use similarities in citations in the future and past respectively, and thus provide a separation naturally (Weinberg, 1974). The results of this type of analysis depends on the preprocessing step of constructing the data vectors and similarities, and great care is needed in interpreting the results (Boyack et al., 2005; Gläser, Glänzel, & Scharnhorst, 2017; van Eck & Waltman, 2009).

The bibliometric data can also be analysed by constructing networks—such as the citation network between journals—and directly finding structure in them using the general purpose tools for analysing the networks. The development of such methods within network science has exploded since massive amounts of data on large variety of networks—such as on social and transportation networks—have become available (Boccaletti, Latora, Moreno, Chavez, & Hwang, 2006; Newman, 2003). A prominent way of finding structure in citation networks using these methods is to investigate network clusters or communities (Fortunato, 2010; Fortunato & Hric, 2016; Porter, Onnela, & Mucha, 2009), which are subnetworks that have a large number of links inside them (Chen & Redner, 2010; Lambiotte & Panzarasa, 2009; Lancichinetti & Fortunato, 2012; Radicchi, Fortunato, & Vespignani, 2012; Rosvall & Bergstrom, 2008). The assumption with most of these methods is that the network is constructed from densely connected cores of nodes or journals that have a relatively small number of citations to the rest of the network. This is in contrast to the methods based on similarity of journals that can find groups with a strong preference for receiving or giving citations from a certain subset of journals, for instance work of applied research can cite theoretical works, without being cited back.

Even if one would accept the premise that the community-like structures are relevant in citation networks, many community detection methods are besieged with intrinsic problems. Very often they detect structures even in case of random networks by mistaking noise for data, they might be very sensitive to small perturbations (noise), and posses a “resolution limit”, i.e. suffering from the inability to identify communities below a certain size that depends on the total size of the network (Fortunato & Barthélemy, 2007; Guimerà, Sales-Pardo, & Amaral, 2004). The performance, reliability, and even the results to some extent depend on the choice of a method from the large set of currently available methods.

The problems with community detection methods are well-known in the network science literature, and the need to find the richer structure in networks than those obtained by partitioning nodes to communities has been acknowledged for many types of networks (Leskovec, Lang, Dasgupta, & Mahoney, 2009; Palla, Derényi, Farkas, & Vicsek, 2005; Rombach, Porter, Fowler, & Mucha, 2014; Wang & Hopcroft, 2010; Xie, Kelley, & Szymanski, 2013). Very recently, as a solution to this problem, the old idea of using stochastic block models (SBMs) as models of network structure (Holland, Laskey, & Leinhardt, 1983; Lorrain & White, 1971; Wasserman & Anderson, 1987) has received renewed attention, because of the theoretical and algorithmic advances that enabled their use in a reliable and scalable way (Bianconi, 2009; Karrer & Newman, 2011; Peixoto, 2012a). SBM is a model in which nodes belong to *blocks* (the name for groups in the SBM paradigm) and edges are created between (and within) the blocks with some fixed probabilities for each pairs of blocks. The methods based on SBMs work by finding the model which best explains the network data. The best explanation is not necessarily the model that would have most likely produced the data, but the simplicity of the model must also be taken into account, and the principled and powerful ideas from statistical inference literature are used to avoid such overfitting. One can consider the blocks as “super nodes” that are connected with weighted edges, and SBM methods then—by definition—try to find the “super network” that is the best simplification of the original network.