A new complex network-based approach to topic modeling

July 30 2018, by Ingrid Fadelli

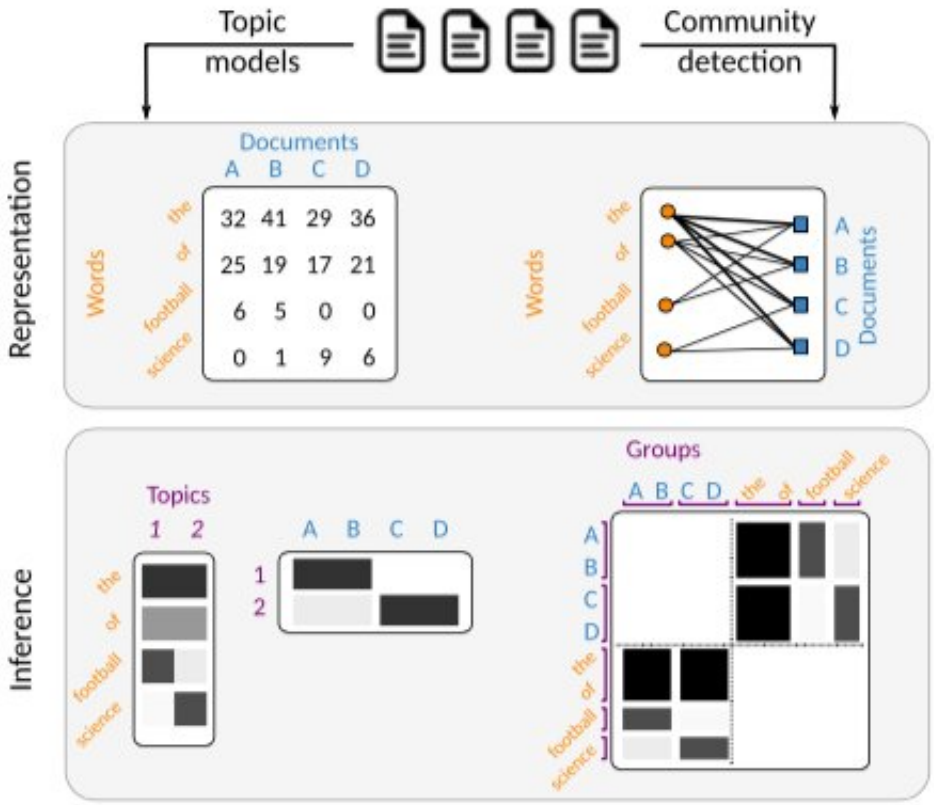

Fig. 1. Two approaches to extract information from collections of texts. Topic models represent the texts as a document-word matrix (how often each word appears in each document), which is then written as a product of two matrices of smaller dimensions with the help of the latent variable topic. The approach we propose here represents texts as a network and infers communities in this network. The nodes consists of documents and words, and the strength of the edge between them is given by the number of occurrences of the word in the document, yielding a bipartite multigraph that is equivalent to the word-document matrix used in topic models.

Credit: Gerlach et al.

Researchers at Northwestern University, the University of Bath, and the University of Sydney have developed a new network approach to topic models, machine learning strategies that can discover abstract topics and semantic structures within text documents.

"One of the main computational and scientific challenges in the modern age is to extract useful information from unstructured texts," the researchers explained in their study. "Topic models are one popular machine-learning approach that infers the latent topical structure of a collection of documents."

Topic models are currently being used to identify semantically related texts and classify documents within a number of fields, including sociology, history, linguistics, and psychology. The most commonly used method, latent Dirichlet allocation (LDA), is also used for bibliometrical, psychological and political analysis, as well as for image processing.

Despite its widespread success, LDA presents several flaws in the way it represents text, such as a lack of method to choose the number of topics, discrepancies with statistical properties of real texts and a lack of justification for the Bayesian prior, which in Bayesian statistical inference is the probability distribution expressed before evidence is presented.

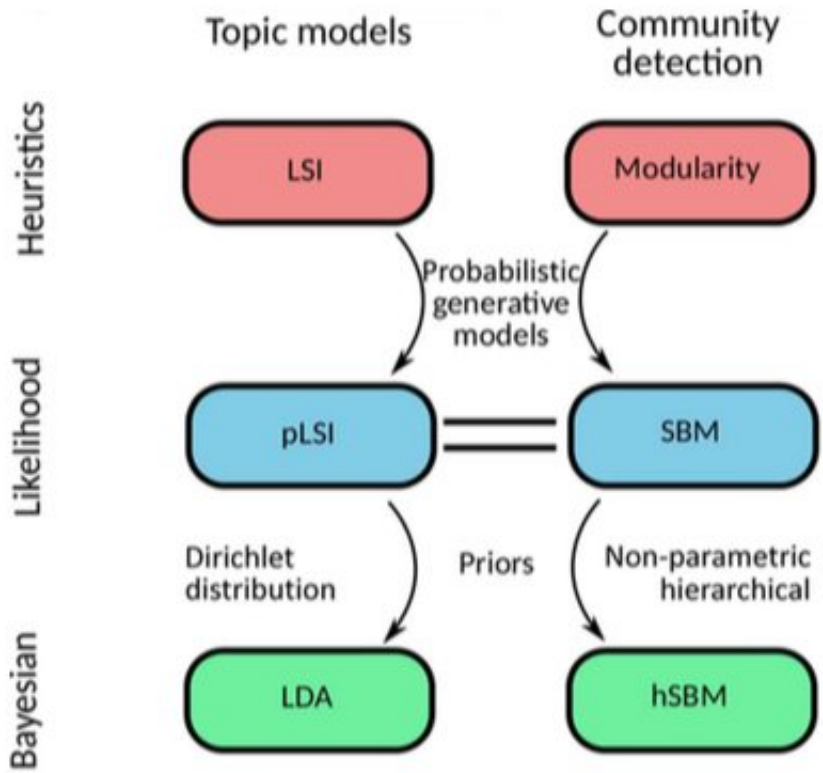

Fig. 2. Parallelism between topic models and community detection methods. The pLSI and SBMs are mathematically equivalent, and therefore, methods from community detection (for example, the hSBM we propose in this study) can be used as alternatives to traditional topic models (for example, LDA).

Credit: Gerlach et al.

A large portion of recent research into topic models has focused on creating more sophisticated versions of LDA that perform better or can effectively analyze particular aspects of documents.

The approach developed by this team of researchers stems from network theory, a theory used in physics and other scientific fields that provides techniques for analyzing graphs, as well as structures in systems with different interacting agents. Their new framework for topic modeling is

based on the approach used to find communities in complex networks, which, in the context of network theory, is a graph with features that occur in modeling of real-life systems.

"I was working on natural language and topic modeling from the perspective of complex systems and complex networks," Martin Gerlach, postdoctoral fellow at Northwestern University told TechXplore. "The problems seemed very similar, yet the communities of computer science (topic modeling) and complex networks seemed to work largely independently. Being trained as a physicist, we wanted to show that two seemingly different problems could be reduced to the same underlying math."

Gerlach and his colleagues devised a new approach to identifying topical structures that relates to the problem of finding communities in complex networks. Their technique represents text corpora as bipartite networks, a class of complex networks that divide nodes into sets X and Y, only allowing connections between nodes in different sets.

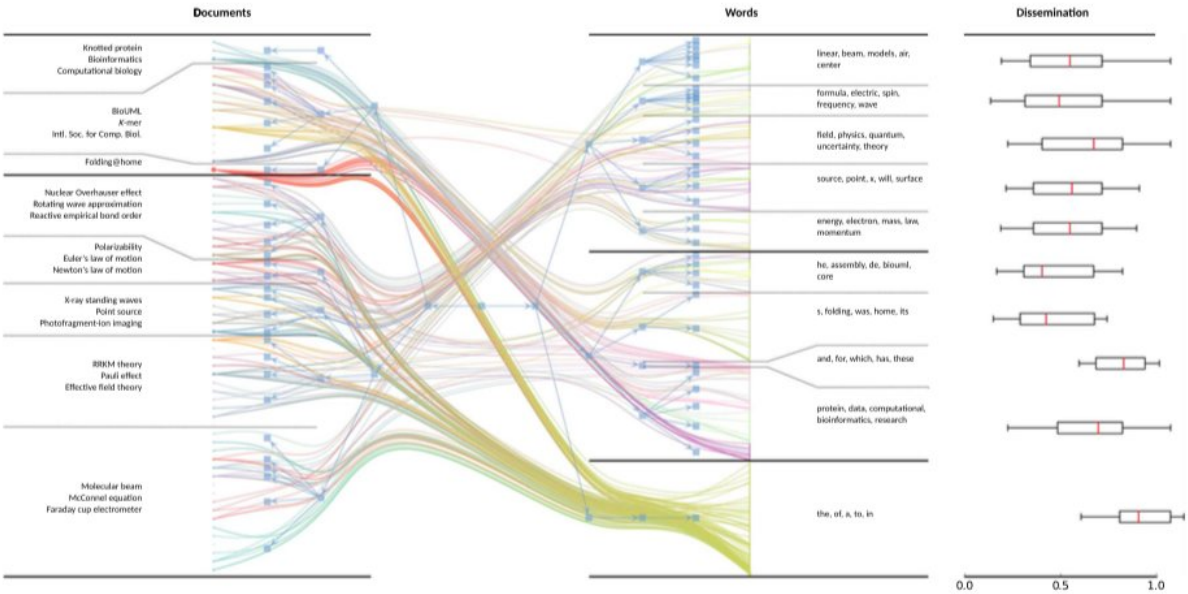

Fig. 5. Inference of hSBM to articles from the Wikipedia. Articles from three categories (chemical physics, experimental physics, and computational biology). The first hierarchical level reflects bipartite nature of the network with document nodes (left) and word nodes (right). The grouping on the second hierarchical level is indicated by solid lines. We show examples for nodes that belong to each group on the third hierarchical level (indicated by dotted lines): For word nodes, we show the five most frequent words; for document nodes, we show three (or fewer) randomly selected articles. For each word, we calculate the dissemination coefficient U_D, which quantifies how unevenly words are distributed among documents (*60*): $U_D = 1$ indicates the expected dissemination from a random null model; the smaller U_D ($0 < U_D < 1$), the more unevenly a word is distributed. We show the 5th, 25th, 50th, 75th, and 95th percentile for each group of word nodes on the third level of the hierarchy. Intl. Soc. for Comp. Biol., International Society for Computational Biology; RRKM theory, Rice-Ramsperger-Kassel-Marcus theory.

Credit: Gerlach et al.

"We mapped the problem of topic modeling to the problem of community detection in a network consisting of words and documents showing that they are mathematically equivalent," explained Gerlach.

The researchers' approach, which adapts existing community-detection methods, was found to be more versatile and principled than other existing topic models, for instance detecting the number of topics present in texts and hierarchically grouping both words and documents. Their method used a stochastic block model (SBM), a generative model for graphs that generally maps communities, subsets of items that are connected with one another.

"We solve some of the intrinsic and known problems of popular topic modeling algorithms such as LDA (e.g. how to determine the number of topics)," said Gerlach. "In addition, our work shows how to formally relate methods from community detection and topic modeling, opening the possibility of cross-fertilization between these two fields."

The SBM approach developed by Gerlach and his colleagues could have interesting applications in other areas where machine learning is used, such as the analysis of genetic codes or images. In future, the researchers plan to continue exploring the potential of complex networks both within the context of text analysis and beyond.

"The equivalence between topic modeling and community detection allows to use insights gained in each of the communities and apply to the other domain," said Gerlach. "I hope to use these insights to gain a better understanding of these machine learning algorithms; why they work, and more importantly, under which conditions they do not work."

More information: A network approach to topic models, Martin Gerlach et al. A network approach to topic models, *Science Advances* (2018). DOI: 10.1126/sciadv.aaq1360