Communities of words

M. Gerlach[1], T. P. Peixoto[2], and E. G. Altmann[1]

(1) Max Planck Institute for the Physics of Complex Systems, Dresden, Germany
(2) Institut für Theoretische Physik, Universität Bremen, Bremen, Germany.

One of the computational and scientific challenges in the age of Internet is to extract useful information contained in massive amounts of written documents. One popular approach to account for the heterogeneous usage of words in documents are *topic models*, in particular its most well-known variant called Latent Dirichlet Allocation (LDA). These methods originated in Machine Learning and are designed to extract the *latent* topical composition from a diverse group of documents. In this work we use tools from statistical physics and complex systems to investigate topic models. We first show that the topicality of texts is consistent with scaling laws of word frequencies (e.g., Zipf's law) but that fluctuations around these laws are much larger than expected from simple null models [1]. We then formulate an alternative approach to topical models by considering the bipartite network composed of documents and words. In this framework, the problem of inferring the underlying topical structure is equivalent to the task of *community detection* in networks. Using *stochastic block models* (SBM) as generative models of network structure [2], we obtain a generalization of LDA. More generally, our work elucidates the rich connection between topic modeling in language and community detection in complex networks [3].

[1] M. Gerlach and E.G. Altmann: "Scaling laws and fluctuations in the statistics of word frequencies" *New J. Phys.* **15**, 113010 (2014).

[2] T.P. Peixoto: "Model selection and hypothesis testing for large-scale network models with overlapping groups" *Phys. Rev. X* **5** 011033 (2015)

[3] M. Gerlach, T.P. Peixoto, and E.G. Altmann: "Stochastic block models and topic modelling" *in preparation*