KNOWLEDGE NETWORKS IN SCIENCE AND TECHNOLOGY

NETSCI 2017 SATELLITE (June 20 2017, Indianapolis, Indiana)

Visualizing Networks of Knowledge
Katy Börner, Indiana University
2:00 - 2:30, Room: White River 106

In the information age, the ability to read and make data visualizations is as important as the ability to read and write. This talk explains and exemplifies the power of data visualizations not only to help locate us in physical space but also to help us understand the extent and structure of our collective knowledge, to identify bursts of activity, pathways of ideas, and borders that beg to be crossed. It introduces historical and novel visualizations of knowledge networks together with a theoretical visualization framework meant to empower anyone to systematically render data into insights. Materials from the Information Visualization MOOC (http://ivmooc.cns.iu.edu) and maps from the Places & Spaces: Mapping Science exhibit (http://scimaps.org) will be used to illustrate key concepts and to inspire participants to visualize their very own data.

- Börner, Katy. 2015. Atlas of Knowledge: Anyone Can Map. Cambridge, MA: The MIT Press.
- Börner, Katy, and David E. Polley. 2014. Visual Insights: A Practical Guide to Making Sense of Data. Cambridge, MA: The MIT Press.
- Börner, Katy. 2010. Atlas of Science: Visualizing What We Know. The MIT Press.

To catch an open access predator: article features predictive of predatory journals
Daniel Acuna, Syracuse University
2:40 - 3:00, Room: White River 106

In this work, we analyze what factors are predictive of an article being published in a predatory journal. We obtained a highly recognized list of predatory journals known as the Beall's list. It is worth mentioning that this list has been shut down—voluntarily, according to the author. While the Beall's list is not perfect, researchers have found that journals from that list are much more likely to accept fake editors then other journals (Sorokowski, Kulczycki, Sorokowska, & Pisanski, 2017). We matched articles published in those predatory journals to the Microsoft Academic Graph (MAG, Sinha et al. (2015)) and to the PubMed Open Access Subset to obtain citation and full text information. We restricted our analysis to the areas of Immunology and Microbiology, Medicine, Biochemistry, Genetics and Molecular Biology, and Neuroscience. For our analysis, we sampled an equal number of articles from non-predatory journals. Based on the factors computed from this dataset, we hope to find how likely they predict predatory publishing behavior.

Viziometrics: mapping the visual literature
Jevin West, University of Washington
3:00 - 3:30, Room: White River 106

Figures are a primary mode for communicating scientific results, yet little has been done to extract and analyze this information at scale. Most of the work in mining the literature has been on full text, citations, or metadata associate with an article. These visual objects are information dense and complex, but as the saying goes, worth a thousand words. In this talk, I will present methods for extracting and mapping this information in order to ask basic questions around visual communication, knowledge organization, and impact.

Maps of sparse Markov chains efficiently reveal overlapping and hierarchical community structure in citation flows with memory
Martin Rosvall, Umeå University
3:40 - 4:00, Room: White River 106

Classification and organization are cornerstones of science. In science itself, flows of ideas persist in scientific fields that reveal the organization of science. Conventional approaches to identify scientific fields in citation flows ignore that the flow direction often depends on more than a single step, that is, where the flows come from. For example, citation flows from biology journals through a multidisciplinary journal typically return to biology journals. We have designed maps of network flows modeled by sparse Markov chains with several advantages: The sparse Markov chains derived with an efficient lumping algorithm compactly represent network flows with memory and the cross-validated maps reveal hierarchically nested and overlapping communities. Furthermore, the method applies to multi-step pathways of any length. For illustration, based on citation flows through 4.5 million scientific articles in 2 billion citation pathways, we classified 10,000 journals into hierarchically nested research fields that overlap in multidisciplinary journals. With small computational overhead, the quality of the journal classification, measured as the fraction of citation flows that stays within the same research field in the next step, increased from 59 percent for a conventional mapping based on first-order Markov chains to 82 percent. Compared with an established classification by Web of Science, which has a module flow persistence of 44 percent, the data-derived journal classification should form better units of analysis for building recommendation systems, comparing impact across research fields, or measuring interdisciplinarity in science of science studies.

Principled Knowledge Discovery from Networks
Tim Weninger, University of Notre Dame
4:00 - 4:30, Room: White River 106

In this talk we present current and ongoing work about how to learn the Lego-like building blocks of real world networks in order to gain insights into the mechanisms that underlie network growth and evolution. We recently discovered a relationship between graph theory and formal language theory that thinks like a context free grammar (CFG), but for graphs. The extracted hyperedge replacement grammar (HRG) contains the precise building blocks of the network as well as the instructions by which these building blocks ought to be pieced together to make predictions about the data. In a second project, we present work on how humans create networks of information and how we can leverage some of those networks pursuit of knowledge. We find that these knowledge networks are typically constructed with a subject, predicate and an object, where the subject and the object typically represent some real-world entity or idea, and the predicate represents a one-way relationship between the subject and the object. Tasks like automated fact checking are easy when a true fact-triple is present in the knowledge graph; however, if a knowledge graph is a missing some fact it could be that (1) the statement is false or (2) the statement is true, but missing from the knowledge graph. In order to rate the validity of some missing fact we have created a system that automatically interprets whether the assertion is true or not.

From zero to one: Success and Innovation
Yi Bu and Ying Ding, Indiana University
4:30 - 5:00, Room: White River 106

Innovation has long been discussed as it helps succeed. However, success is not easy, especially the step from "zero" (non-existence) to "one" (existence). In this talk, we will first showcase three different levels of innovations (incremental, breakthrough, and transformational innovations) and then present three research questions in the field of science of science: (1) How difficult to leap from zero to one? (2) Should we stand on the shoulder of giants? (3) Will persistent help? We find that beginnings are often difficult, and standing on the shoulder of giants can help young scholars achieve higher but it is not always the wisest way. We also find that persistent publication flows make a more successful career but persistent collaborations do not.

Do patent citations predict changes in the rate of technology improvement?
Jeff Alstott, Massachusetts Institute of Technology
James Savage, University of Virginia
5:00 - 5:20, Room: White River 106

Technologies, from computer chips to solar power, improve over time in apparent trends. The speeds of those trends have a signal in patents and their citations: technologies with faster improvements have patents that cite more central patents [1]. This correlation possibly indicates that technologies are improved more rapidly by leverage existing knowledge more effectively. However, technologies frequently deviate from their average trends, and predicting future technology development using trend extrapolation is more accurate when it accounts for a technology's variability [2]. Can variability in the patent citation record partially account for variability in technology development trends? We examined citation data from over 5 million patents and performance data from 21 technology domains with average annual improvement speeds ranging from ~0% to 85%. We used Bayesian inference to model the technology time series as a function of both trend-following and patent citations. We find that adding patent citation data on top of a simple trend-following model improves the model's fit, though perhaps only weakly. We are using this data to build a predictive model, with the aim of predicting technology performance with greater accuracy than trend extrapolation.

[1] G. Triulzi and C. Magee, "Predicting Technology Performance by Mining Patent Data." Global TechMining Conference (2016).
[2] J. D. Farmer and F. Lafond, "How Predictable Is Technological Progress?" Research Policy 45(3):647-65 (2016).

Small Teams Generate New Directions in Science and Technology; their Decline may Slow Advance
Lingfei Wu, University of Chicago
5:20 - 5:40, Room: White River 106

Teams of growing size have become ubiquitous in the production of high-impact science and technology. Research on memory, idea generation, decision-making, and efficiency suggests that teams of different sizes may not be equal. Here we investigate whether small and large teams differ in the type of advances they contribute. Analyzing teamwork represented by tens of millions of research articles, patents, and software, we find evidence that small teams disrupt the system by drawing on older and less popular ideas, while larger teams pursue popular ideas from the immediate past. As a result, attention to large teamwork comes immediately, while smaller team advances are more often ahead of their time. We relate this difference to the economy of science and technology by showing how grants favor large, risk-averse teams, and we show how the decline of small teams may pose challenges for the long-term vitality of science and technology.

Clustering conference abstracts using a combination of author preferences and topic relevance

Titipat Achakulvisut, Northwestern University

5:40 - 6:00, Room: White River 106

Scientific conferences are the prime venue where new scientific ideas are presented. For large conferences, determining the proper physical location of posters often rely on authors providing keywords and topics. Unfortunately, this is time consuming, sometimes not provided and also may not capture the changing topics covered by cutting edge science. Ideally, we would like to have semi-automated methods that tradeoff the author location preferences with the actual content of the posters.

In this work, we develop a method to automatically combine human provided topics and poster abstracts to automatically collocate posters. We apply this method to the largest Neuroscience conference in the world, Neuroscience Conference in 2016, with approximately 14,000 abstracts. The method relies on building a network of papers that are connected by two types of links: a content distance link and a human topic distance link. The algorithm then applies a hierarchical clustering technique to assign locations to posters. We found that using the abstract with given topic proposed authors can substantially bring down consistence of overall topic.

This method has multiple additional applications such as reviewer-manuscript assignment. In the future, we will extend to cases in which there could be multiple preferences such as conference organizers, chairs, and authors.