What happens when information gets to be free? Exploring temporal co-citation proximity in the Sci-Hub network

Overview

A core aim of scientometrics is to provide a quantitative description of the structure of scientific activity. One of the central methods employed is the creation and analysis of maps of scientific activity, with the data underlying most of these representations derived from published scientific literature. By creating graphs of author collaborations or co-citations, we are able to gain a better understanding of disciplinary relationships and learn about the ways that science is created. However, we know less about the way that science is consumed. This is due to the fact that usage data (e.g., number of downloads of a particular article over time) is hard to come by, and often provided in aggregate, without geolocation data or timestamps. None of these limitations apply to the usage data for Sci-Hub, the largest online repository of pirated academic literature. Here, we propose a number of different analyses of the Sci-Hub data, ranging from simple descriptions of the network structure to novel techniques that leverage the data's unique temporal dynamics.

Related Work

The three ideas we wanted to learn more about before starting the project were (1) previous attempts to map scientific activity, (2) ways of modeling networks like Sci-Hub, and (3) ways of modeling evolving networks.

Design and Update of a Classification System: The UCSD Map of Science (Börner et al., 2012)

As of 2012, there are over 200 maps of scientific activity. All of these take a unique approach to reducing a high-dimensional dataset down to two dimensions. The authors of this paper created one of the first widely-used map of science, and in this paper they update it with new data. A comparison of the 5-year and (updated) 10-year maps shows an increase in the number of journals that can be mapped and a more even distribution of journals across disciplines. The actual methods used to create the map are not listed in the article, and even in the supplementary material the exact procedures are unclear. Interestingly, as an intermediate step the authors created 18 different network representations for the same set of journals (frequency

matrices) and then combined them into a single representation. This will be something to look into more carefully when creating our graph representation of the Sci-Hub dataset.

Without a clear description of the network construction, it is difficult to understand the tradeoffs that the authors made when constructing their frequency matrices (i.e., network representations), and unclear how their decisions may have affected the resulting map of science. One dubious decision mentioned in the paper was to throw out interdisciplinary journals because they defied classification; but the authors may have inadvertently thrown out important bridge-nodes and destroyed important network structure. This paper showed us that the construction of our network representation should be carefully considered, and that it is not nearly as straightforward as we previously believed. The supplementary material from this paper and prior, related research will help us in that regard.

Co-Citation in the Scientific Literature: A new measure of relationship between two documents (Small, 1973)

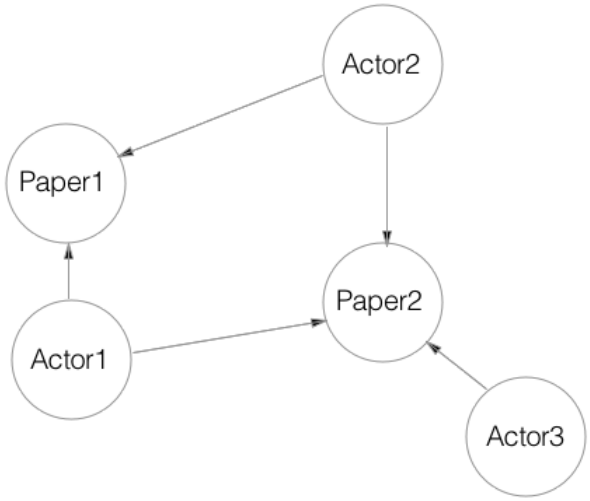
This is one of the foundational papers in the field of bibliometrics, and crucial for us considering the dataset that we are working with. Co-citation is defined as “the frequency with which two items of earlier literature are cited together by the later literature” (Small, 1973). The number of times that two papers are cited together gives the co-citation index for that paper. Unlike bibliographic coupling, which is static and depends only on the shared citations contained within two documents, co-citation is dynamic and changes along with the important topics of a field. Not only that, but bibliographic coupling between two papers often diverges from their co-citation index, with the authors arguing that co-citation is a better index of subject similarity.

(Note that one of the ideas we have for structuring the Sci-Hub data is inspired by co-citation. We plan to create a co-download network, with each node representing a journal article (or perhaps a journal) and edges between nodes representing the number of times that two articles/journals were downloaded together. There is more on this in the project proposal below.)

The authors compare co-citation to two other similar measures (bibliographic coupling and direct citation), but do not provide any further analysis of the network. They do not generate comparison networks according to their theory of subject similarity, nor do they compare the co-citation network to a null model. Nevertheless, the idea of co-citation provided a useful jumping off point for our project, and we plan to read more contemporary literature on the topic to help us with our analyses.

A theory of fads, fashion, custom, and cultural change as informational cascades (Bikhchandani, Hirshleifer, & Welch, 1992)

We originally chose to read this paper because it seemed to provide a good starting point for understanding, modeling, and analyzing network changes over time. Indeed, when a network consists of actors who are capable of observing the actions of their neighbors, informational cascades are an important concept to consider. However, while reading this paper we had an important realization: when modeling the Sci-Hub network as a co-download graph, the “actors” in our graph are incapable of perceiving the actions of those around them. Thus, we should not expect to see informational cascades. That is because even if multiple people download the same article, the resulting directed graph structure does not allow for information to flow between individuals. In the example graph below, notice how the graph can be broken into many weakly connected components, with each containing only a single actor and a collection of papers.



Nevertheless, we still plan to look for cascading behavior in the graph. The authors describe how cascades can be affected by the release of public information, and any cascades that we observe in the Sci-Hub network may be related to press releases, news stories, or major scientific breakthroughs.

A major weakness of this paper is the lack of any analysis of empirical data. After going into the mathematical properties of informational cascades in great detail, the authors do not construct a model network, nor do they analyze an existing network. Nevertheless, the section on the effects of public information on information cascades provides an interesting analysis possibility for our data.

Project Proposal

Introduction

The objective of scientometrics is to describe science as a measurable entity (Price 1986: xvi). At the core of this endeavor is the decomposition of scientific literature into disciplinary and sub-disciplinary structures (Leydersoff 2009: 348). A map of science gives us valuable insight into flows of knowledge, sites of interdisciplinary collaboration, areas of consensus and structural holes between domains (Börner et al., 2012).

In recent years, usage data has allowed researchers to map science as it presently occurs in the varied, unsanctioned inquiries of a multitude of students, experts and laymen. With usage data, scientometrics has gained greater granularity as well as a means to overcome citation delay and peer-review filters, among other hindrances to mapping posed by the publication industry (Bollen and Van De Sompel 2006: 228). Despite the fruitfulness of this new venue of scientometrics, researchers have had to face the uneven quality and coverage of usage data.

Publishers and libraries provide usage data without geolocation or granular timestamps. Moreover, this datasets are usually based on aggregated user-level observations so that it is impossible to follow readers over time. Even if publishers were to provide detailed data, researchers would fail to represent usage in the global scientific community since a large portion of this community lacks legal access to the large repositories of literature.

None of the limitations described above apply to the usage data for Sci-Hub, the largest pirate repository of scholarly literature. For all its affordances (see Data section), Sci-Hub poses the thorny issue of how to map science with great temporal and geographical granularity in a meaningful way.

In order to meet this challenge we will ask, what structural attributes of science can we capture if we allow for users to connect publications in increasingly longer tasks? If we imagine a student searching for references to write a blog post, a final paper, or a dissertation, we can assume that she will invest a different amount of time for each task, then at what degree of time investment do we expect interdisciplinary or specialized connections to emerge, and what does this mean for the general structure of science? Do these attributes change from one country to another? Would we expect to find these attributes at random? To answer this questions we will adapt co-citation proximity analysis to download histories and employ it to describe and compare network attributes of science across countries.

Data

We will use the download logs from the research database Sci-Hub for the period September 1 2015 through February 29 2016 (Elbakyan and Bohannon, 2016). The logs consist of 28 million download request events from more than 3 million users. The entries are time-stamped and geolocated and the user IP addresses have been replaced by anonymized unique identifiers. In addition, we have collected metadata from Crossref for the 10 million papers requested by Sci-Hub users in this period. These metadata include title, author, journal, issue date, subject and author affiliation when available.

Methods

Co-citation analysis quantifies the relation between two documents that co-occur as bibliographic references in other papers (Liu and Chen 2011: 2). A co-citation graph can be constructed from a similarity matrix where the cells represent counts of co-occurrence (Leydersoff 2005). Larger counts are recognized in the literature as an indicator of stronger thematic relationship between articles. Analogously, we will posit that the relation between two articles can be described by the counts of user download histories in which they co-occur. Furthermore we will aggregate these relations to a larger bibliographical unit, journals.

Journals have been the basic unit for mapping science for over 45 years (Boyack 2005: 461). Scientometric papers frequently discuss aggregate journal to journal co-citation graphs as “operational indicator[s] for the discipline organization of the sciences” (Hu 2011:658). This is because, as Leydersoff et al. concluded in a review of recent efforts to map science, “journals group naturally in the network of aggregated citation relations, and thus shape an ecology” (Leydersoff 2015: 1001). Journal-journal is also the scale at which Bollen and Van De Sompel constructed their map of science based on usage data collected for the Los Alamos National Laboratory research community (2006). What we will add to their framework, is a notion of how proximity of downloads impacts our ability to record the similarity of journals.

The proximity at which two references occur in a paper may be an important factor in determining their thematic similarity (Gipp 2009). Documents referenced in the same sentence, paragraph or section may be fulfilling similar or complementary functions within that citation context. An analogous reasoning leads us to believe that papers downloaded within a similar time-frame may be thematically related within the context of a user’s task, such as writing a paper, designing a syllabus, or preparing for a class. Therefore we will adapt what is known as Co-citation proximity analysis (CPA) to quantify the similarity of journal references given their temporal distance in users’ download history.

Proposed Algorithm

We will use the CPA algorithm employed by Boyack and Small (2013: 160) modified for co-download matrices and temporal distances:

1. Set downloading thresholds to obtain subsets of downloaded journal references that can then be grouped into T trial sets.
2. For each trial apply every co-download weighting scheme in W. For each trial and scheme, calculate weighted co-downloads on a per user basis, and convert into modified frequencies where f = wt/log(n*(n+1)/2), wt is the weighting factor and n is the number of downloaded journal references for the user.
3. Sum modified frequencies f by co-download pair over all users. Calculate cosine index similarity values for each pair of journals as $S_{ij} = f_{ij} / sqrt(\sum(f_i)* \sum(f_j))$.
4. Filter these similarity matrices to include only the top-n S_{ij} values for each node i, where n varies in an interval defined by the log of column sums $\sum(f_i)$.

We will define the trial sets in T at the quartile intervals of the journals' download frequency distribution. We will define the weight schemes in W as follows: In the first scheme relative positions will be calculated using minute offsets normalized by the length download history, and transformed into a centile position within the body of said history (Boyack and Small 2013: 158). In the second scheme, every co-occurrence of journals in a download history will receive a value of $1/2^n$ where n is the time-frame for the co-download (Gipp 2009: 3).

Time-frame	**n**
Hour	0
Day	1
Week	2
Month	3
Quarter	4
Semester	5

The output will be W*T similarity matrices from which we will construct a series of journal co-download graphs. For each graph we will calculate the relationship between a series of measurements and co-download proximity. The measures are:

Measure	Name	Definition
1	Co-download frequency	Total number of edges
2	Consensus link frequency	Number of paired relationships that occurred in at least in 50% of the countries (Boyack 2005: 466)
3	Quadratic entropy or Rao-Stirling Interdisciplinarity	$\sum_{ij}(p_i p_j)d_{ij}$ where p_i and p_j are proportional representations of the journals i and j in the system and d_{ij} is the degree of difference (disparity) between journals i and j (Leydersoff 2015: 1004)
4	Average Clustering	Average of clustering coefficient over every journal
5	Density	Number of edges over number of possible edges

Statistical Analysis

Based on Shi et al.'s analysis of (2010: 53) we propose the following testing framework: For every journal co-download graph G_j we will construct a random graph G_r with the same degree sequence. This means that for every trial-scheme combination that produces a G_j there is a corresponding graph Gr with the same number of nodes and same number of edges. In every G_j, G_r pair, the journals will have the same degree (number of neighbouring journals). From this it follows that the density and the degree sequence of G_r are exactly the same as those of G_j. We will compare every G_j and G_r on the measures presented in the previous section. We will evaluate our success by our ability to capture structural attributes of the discipline organization of the sciences that are significantly distinct from randomly emerging attributes.

BIBLIOGRAPHY

Bohannon J (2016) Who's downloading pirated papers? Everyone. Science 352(6285): 508-512.http://dx.doi.org/10.1126/science.352.6285.508

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., … Boyack, K. W. (2012). Design and Update of a Classification System: The UCSD Map of Science. PLOS ONE, 7(7), e39464. https://doi.org/10.1371/journal.pone.0039464

Boyack, K. W., Small, H. and Klavans, R. (2013), Improving the accuracy of co-citation clustering using full text. J Am Soc Inf Sci Tec, 64: 1759–1767. doi:10.1002/asi.22896

Bollen, J. and Van De Sompel, H. Mapping the structure of science through usage Scientometrics, Vol. 69, No. 2 (2006) 227–258

Elbakyan A, Bohannon J (2016) Data from: Who's downloading pirated papers? Everyone. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.q447c

Gipp B. and Beel, J. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In B. Larsen and J. Leta, editors, Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935.

Hu, C., Hu, J., Gao, Y. et al. Scientometrics (2011) 86: 657. doi:10.1007/s11192-010-0313-6

Klavans, R. and Boyack, K. W. (2009), Toward a consensus map of science. J. Am. Soc. Inf. Sci., 60: 455–476. doi:10.1002/asi.20991

Leydersoff, L. Co-occurrence Matrices and their Applications in Information Science: Extending ACA to the Web Environment Journal of the American Society for Information Science and Technology

Leydesdorff, L., de Moya-Anegón, F. and Guerrero-Bote, V. P. (2015), Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of Scopus data (1996–2012). J Assn Inf Sci Tec, 66: 1001–1016. doi:10.1002/asi.23243

Leydesdorff, L. and Rafols, I. (2009), A global map of science based on the ISI subject categories. J. Am. Soc. Inf. Sci., 60: 348–362. doi:10.1002/asi.20967

Shi, X., Leskovec, J., McFarland, D. Citing for High Impact JCDL '10 Proceedings of the 10th annual joint conference on Digital libraries Pages 49-58

Liu, S. and Chen, C. The Effects of Co-citation Proximity on Co-citation Analysis The 13th Conference of the International Society for Scientometrics and Informetrics (ISSI), July 4-7, 2011 Durban, South Africa.

Price, Derek J. de Solla Little science, big science—and beyond. New York: Columbia University Press 1986

Tsay, M., Xu, H. & Wu, C. Scientometrics (2003) 57: 7. doi:10.1023/A:1023667318934