Slow science, fast science. Exploring temporal proximity in Sci-Hub's co-download network

Abstract

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 to explore the evolution of disciplinary boundaries.

Introduction

In recent years, usage data has allowed researchers to map science through the queries of a multitude of students, experts and laymen. As of 2012, there are over 200 maps of scientific activity (Börner et al., 2012). With usage data, scientometrics has gained greater granularity as well as a means to overcome citation delay and peer-review filters (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, these data 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 main repositories of academic literature. The usage data from Sci-Hub overcomes all these conventional limitations. This makes these novel data the ideal source to analyze the temporal dynamics of information seeking in a global community of academic literature users.

Our research on the temporal dynamics of disciplinary boundaries is intended as a contribution to the current research program in the Information Systems community regarding the conditions for successful boundary spanning. Within this agenda disciplinary boundaries are

understood as those that enclose distinct forms of knowledge production through the enforcing of norms and standards in scholarship (Vashist et al. 2011). The spanning of this boundaries is considered a situated achievement. Drawing from the literature on information seeking behaviour we will argue that a fundamental contextual element in the production of disciplinary boundaries and in boundary spanning is time. We understand temporal dynamics as the temporal patterns and trends in information behavior streams (Whiting 2015: 25). The temporal dynamics we explore in detail here regard the disciplinary boundaries between journals and how Sci-Hub users reproduce or challenge them when seeking information within different time horizons (Savolainen 2006: 112) — that is, in downloading spells that last from a few seconds up to six months. Therefore, we will be addressing the following research question:

I. How do disciplinary ties between journals change with expanding time horizons?
II. How does the emerging network structure of these ties compare to a network based on published scholars' information behavior?
III. What do the methods developed to understand the temporal dynamics of Sci-Hub contribute to the mapping science?

Dataset

We are working with 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. Here we will briefly describe what these data tells us about the site's users and its archive. At the end of this section we will also explain how we constructed and stored a database of download logs with additional metadata.

About 75% of users in Scihub visit the repository only once. As we can see in Figure 1, the number of users that employ the service for more than a day decreases rapidly, but a small spike in the right tail indicates that there is a small group of long-standing users with a few weeks and up to six months of visits.

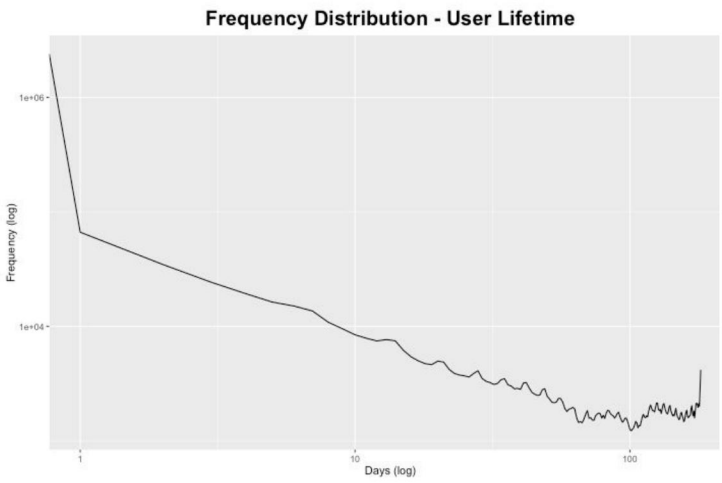


Figure 1. User Lifetime

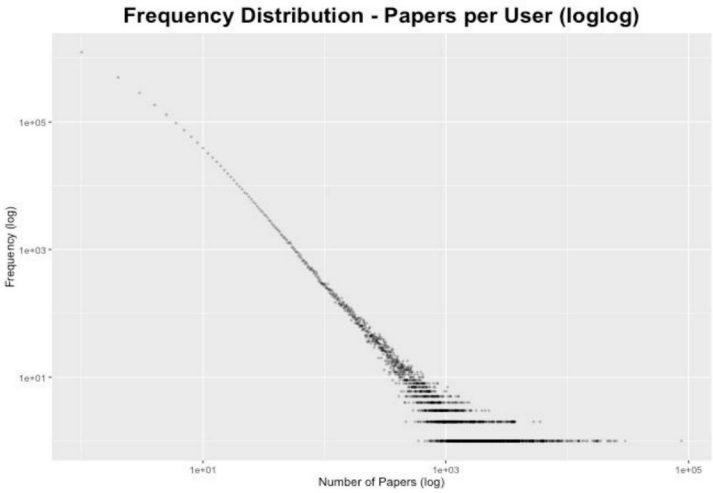


Figure 2. Number of papers per User

Analogously, only 25% of the users download more that 5 articles. Figure 2 shows that larger numbers of downloads per user are infrequent but vary from 1000 papers up to 87000.

Sci-Hub's logs include include download requests for 10.6 million papers across all major domains (See Table 1). About 1.8 million of these requests are malformed and lead to no meta-data and another 1.4 million are not requests for journal articles but for proceedings articles (719 thousand) and book chapters (586 thousand), among other types of documents.

Table 1: Number of Articles by Scientific Domain

Domain	Articles
Applied Sciences	1,536,129
Arts and Humanities	88,001
Economic and Social Sciences	349,647
Health Sciences	2,220,443
Natural Sciences	2,050,401
General	101,350
Total	6,345,971

The download requests further show that users are biased towards more recent literature, with half of all papers belonging to the last decade (See Figure 3). In addition, we found that in 99% of cases, a given paper is requested only in 4% of all countries or less, which points to the the fact that the vast majority of papers don't achieve wide international diffusion. In the case of journals, half of them get more than 9% international coverage and about 10% of journals manage to diffuse across more than 40% of countries (See Figure 4) .

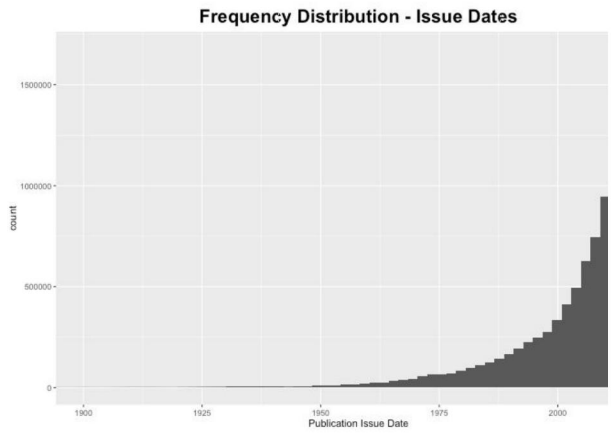


Figure 3. Year Distribution of Articles

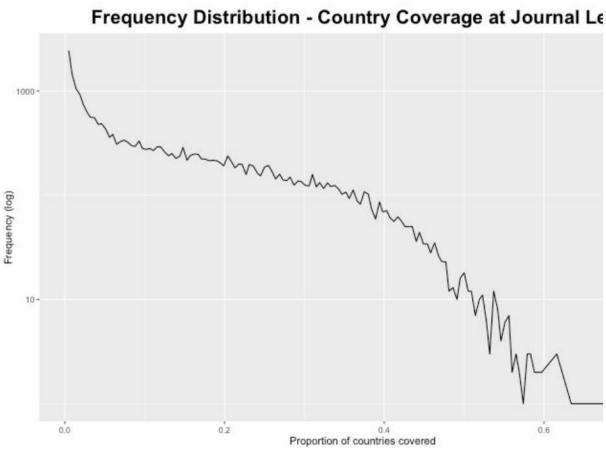


Figure 4. Country Coverage at Journal Level

The download logs were formatted as SQL tables and uploaded to Google BigQuery. We were able to acquire additional metadata for more than 8 million download requests including document title, document type, journal title, issn, subject keywords, and citation count. We also linked 10 thousand out of 26 thousand journals to domain, field, and subfield categories using the Science-Metrix classification (Archambault 2011) . From this data set we constructed a table of 1.3 million users with fully documented downloads and download times. This table constitutes the population frame from which we sampled 10 thousand users. Since we only sample users who download more than a single article, our sample is skewed in favor of users with larger download histories. User lifetime is similarly skewed in the sample in favor of users that stay longer (See Figures Annex).

Methods

Co-citation is defined as “the frequency with which two items of earlier literature are cited together by the later literature” (Small, 1973), this measure 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 similarity of two articles can be captured by counting the number of times that both articles co-occur in the download histories of Sci-Hub users. We refer to this metric as the article’s co-download count. Once we have computed the co-download count for all pairs of articles, we aggregate the data to the journal-level.

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). There are also recent examples of journal level maps of science, as shown by the work of Bollen and Van De Sompel with 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.

Algorithms and Graphs

In our graphs nodes represent journals. Each node has three attributes, domain, field and subfield. These attributes are based on the Science-Metrix journal classification (Archambault 2011) which combines an algorithmic classifier trained on existing journal classifications (ISI, CHI, ERA) and expert opinion[1]. Domain, Field and Subfield are nested categories that range from broad distinctions between the sciences and the humanities (domain), to distinctions between disciplines (field), and finally to boundaries between research areas (subfields). We will use these categories as indices of disciplinary boundaries in our analyses of assortativity and community identification.

An edge occurs between two journals when a user requests downloads from both of them; in the last section we defined this as a co-download. The weight on an edge represents a modified co-download count. The modified count penalizes users with too many download requests. Each user contributes $(log((n(n+1))/2))^{-1}$ to each of the connections between journal that she makes, where n is the size of her download history. This is a conservative way to thin out the influence of heavy downloaders in the aggregate journal-level network. The maximum value for the modified count that a single user can contribute to an edge is 0.91 (when n = 2) and the minimum value is 0.076 (when n = 1000). Note that this upper and lower bound are defined with respect to cut-offs we set for processing users. A minimum of two downloads from different journals is needed to have users build edges, and a maximum of a thousand downloads limits computing time significantly while only excluding 0.1% of the users in the population frame from participating in the graph.

[1] Science-Metrix(2016) Classification http://www.science-metrix.com/en/classification

Creating Subgraphs, Weighting Edges, and Null Models

We are interested in understanding the relationship between co-download time and journal similarity. We hypothesize that if two journals are often downloaded by users in small time windows, this is a reliable indicator of journal similarity. The closer in time two journals are downloaded (on average), the more similar the two journals.

To test this hypothesis we constructed 96 subgraphs containing data from different time windows and computed graph statistics on each subgraph. Time windows were defined in exponentially growing intervals of hours. We created a 96 subgraphs for cumulative time windows starting with the interval [0, 0.001] and up to the interval [0, 3993.9]. To split the full graph into relevant subgraphs we used the CPA algorithm employed by Boyack and Small (2013: 160) modified for co-download matrices and temporal distances. This algorithm works as follows:

1. Set T lower and upper time windows to obtain subsets of downloaded journal references that can then be grouped into T subgraphs. Here we chose a set of exponentially increasing upper time windows $W = \{w_1, w_2, ..., w_k\}$ where $w_k = 0.0001 \cdot 1.2^k$. The lower time window was fixed to zero.
2. For each subgraph, calculate edge weights as follows.
    a. For each user, calculate a modified frequency count $f = (log((n(n+1))/2))^{-1}$, where n is the number of downloaded journal references for the user.
    b. For each edge, sum the modified frequency counts over all users who created that edge. This is the final edge weight.
3. Filter edges to include only the top 90% of edges by weight.
4. For each subgraph, we also created a corresponding null graph using the configuration model. This allows us to create graphs with identical degree distributions but randomized connections.

For each subgraph, we calculated the following statistics for both the Sci-Hub network and the corresponding null network.

- Total edge count
- Total node count
- Degree distribution
- Edge weight distribution
- Largest connected component size
- Clustering coefficient
- Community structure
- Modularity
- Community purity
- Assortativity

Validation

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 using the configuration model. This means that for every time window that produces a G_j there is a corresponding graph G_r 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 compare every G_j and G_r on the measures presented in the previous section. We 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.

For Assortativity we also construct a benchmark graph based on journal-level cocitation. This graph is based on citation metadata from Web of Science for articles from 1990 to 2012. These data are available through the Mimir project in the Graduate School of Education to members of the McFarland Lab. The co-citation network G_c was trimmed at each time window to include exclusively the journals in G_j. This is an artificial way of introduce time into G_c in order to make comparisons possible. A more appropriate comparison would employ a graph based on citation proximity analysis or a graph based on usage from official repositories (See Next Steps).

In addition we validate the use of time proximity by analyzing its relation to linguistic similarity at the journal and article level. For this purpose we construct a tf-idf cosine similarity measure trained on document titles.

Analysis and Results

Following the method outlined in the previous section, we generated 96 cumulative subgraphs. Here we report statistics on these subgraphs which allow us to better understand (1) the relationship between co-download time and journal similarity and (2) the evolution of the Sci-Hub network from a few different perspectives.

Cumulative Edge and Degree Count

In the cumulative Sci-Hub graph, the number of nodes in each graph describes the number of journals that have been downloaded up to that point, while the number of edges describes how many pairs of journals (out of all possible pairs) have been downloaded together. In the plots below, we note that the number of journals in the network increases exponentially up to 0.01 hours (6 minutes) then tapers off at around 10 hours. The number of edges follows a different pattern. It appears that the number of edges increases exponentially up to around 1 hour, then begins to increase at a doubly exponential rate after that. We node that the edge count for the null graph diverges from the Sci-Hub edge count as the number of edges increases above

1000000. This is due to the configuration model algorithm, which creates self-edges that are ignored when we calculate the total edge count.

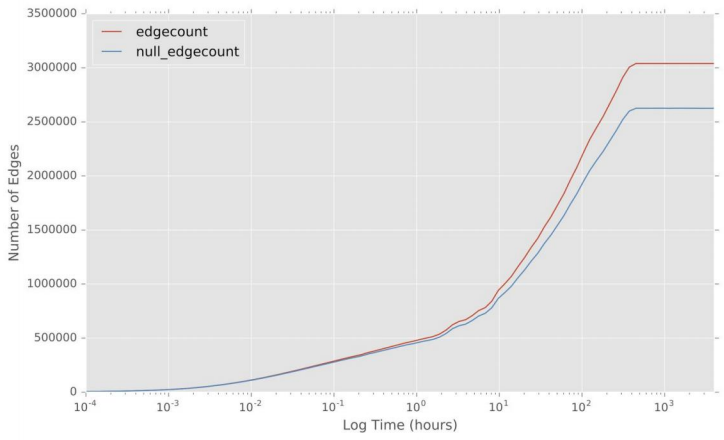


Figure 6. Cumulative Edge Count

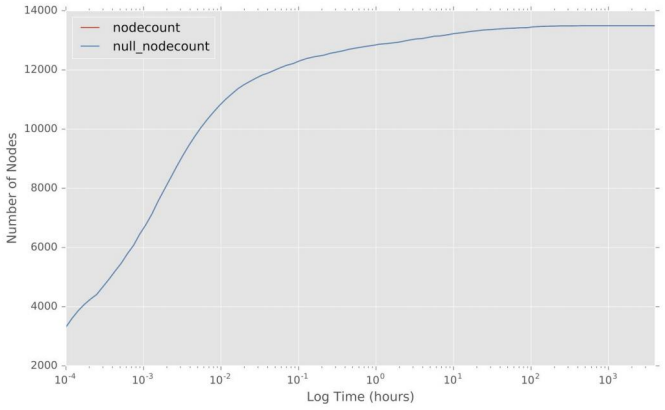


Figure 7. Cumulative Node Count

Cumulative Degree Distribution

In the plot below, we show the degree distribution for each increasing cumulative time window. The degree distribution for the smallest time window is drawn in black, while the degree distribution for the largest time window is drawn in yellow. What we see from the plot is that across all time windows, the degree distribution follows a power law. However, it appears that as the time window increases, alpha decreases.

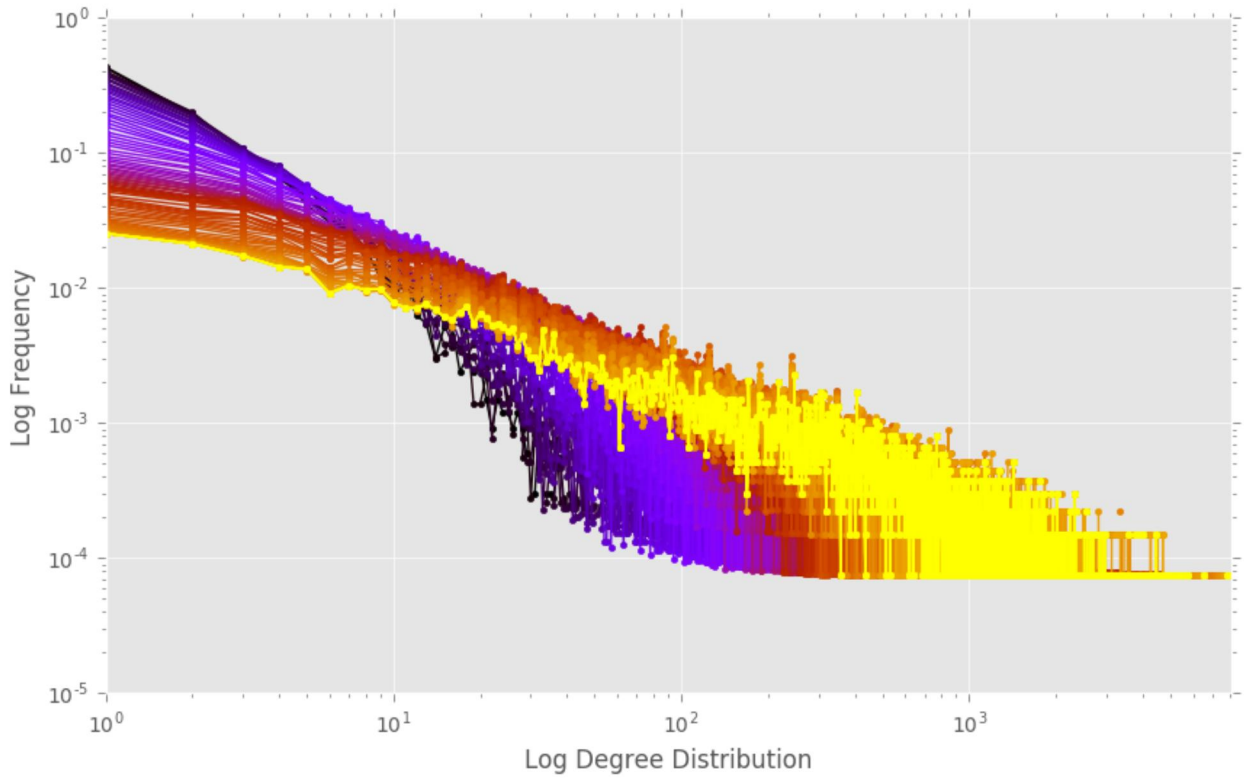


Figure 8. Evolution of degree distribution across time windows. Colors represent size of cumulative time window. Colder colors represent smaller windows.

In the next plot, we see that this is in fact the case. We plot alpha (calculated using maximum likelihood estimation) for each time window and note a clear decrease from 2.3 (smallest time window) to 1.2.

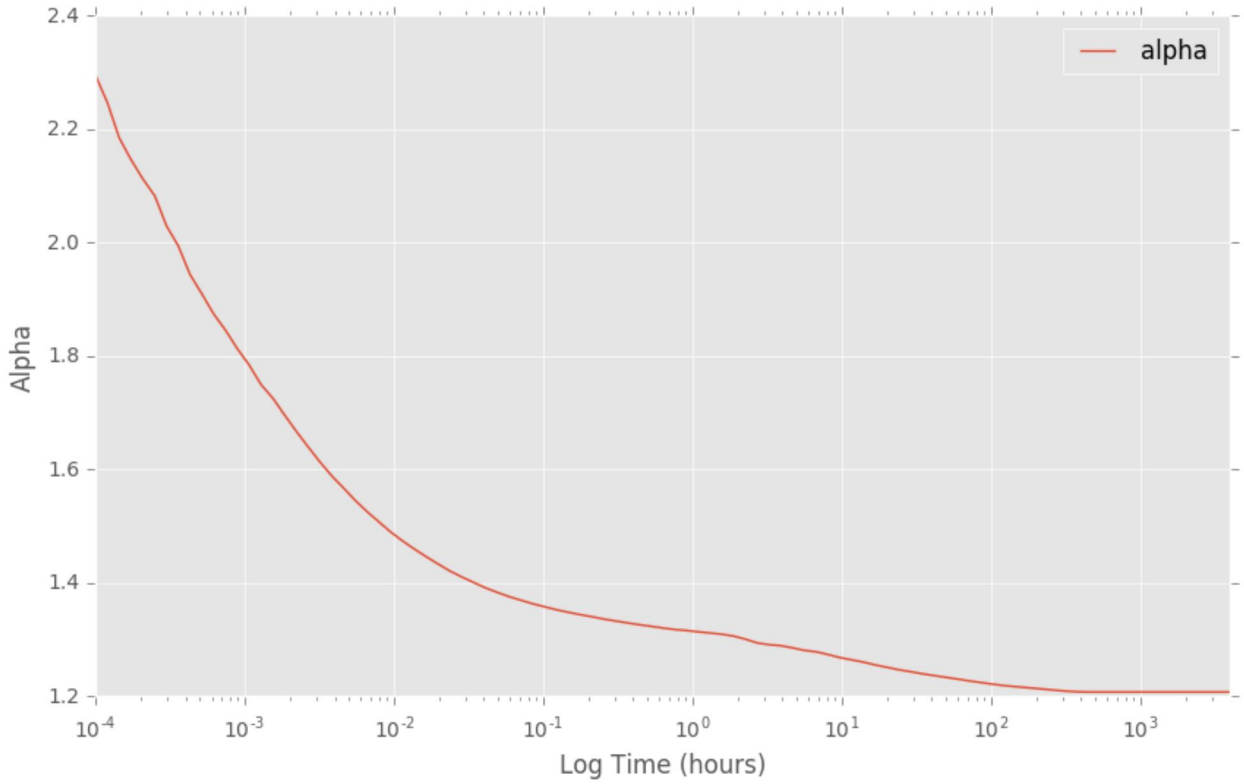


Figure 9. Evolution of Alpha across time windows.

What does this tell us? We see that over very small time periods, the scale-free properties of the network are relevant. However, once the window expands to 36 seconds, alpha drops below 2 and the graph becomes difficult to characterize. On small time scales, the network behaves similarly to a co-citation network. This provides further evidence that there is a strong relationship between co-download time and meaningful structure in the graph.

Cumulative Edge Weight Distribution

The edge weight distribution provides a description of the frequency of edge weights across the entire network. In the graph below, we show how the edge weight distribution changes across different time windows. We note that at small time windows the edge weight distribution is fairly chaotic (black dots), but as the time windows increase the edge weight distribution clearly evolves into a power law distribution (yellow dots).

How can we interpret this? As the time window increases, the graph evolves. Most pairs of journals have weak edges between them, meaning few users download those journals together. However, as the time window increases, the journal pairs which started out with strong connections continue to get stronger, while more and more pairs of journals with weak connections are created. Since this evolves into a Pareto distribution, we can watch over time as the rich get richer; or to put it another way, the most influential journals within disciplines continue to gain in influence.

This is the kind of phenomenon that we normally see in co-citation networks, where authors have full knowledge of the influential papers in the network. It is remarkable to see it at play in the Sci-Hub network, where there is no way for users to see what papers are being downloaded by other users. We posit that while the Sci-Hub network does not inform users of the social dynamics of the scientific community, it does manage to capture the social dynamics of communities of researchers that exist outside the Sci-Hub network.

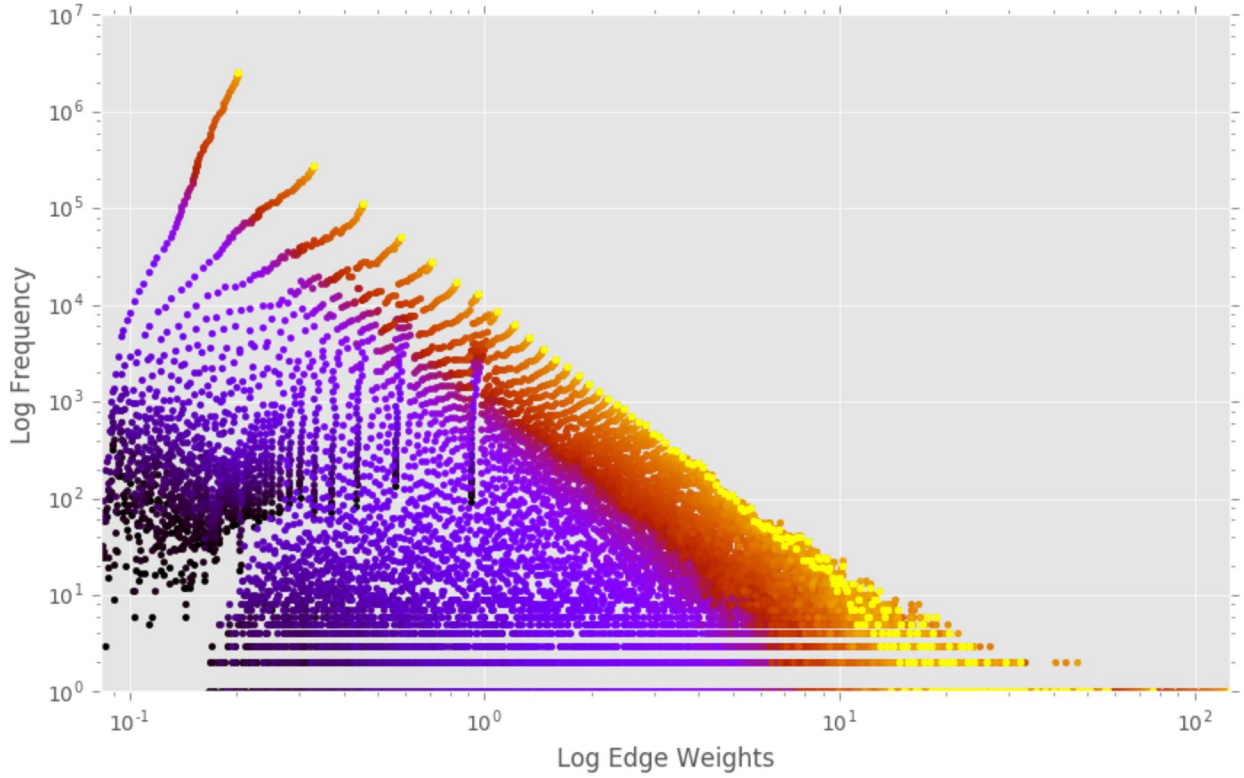
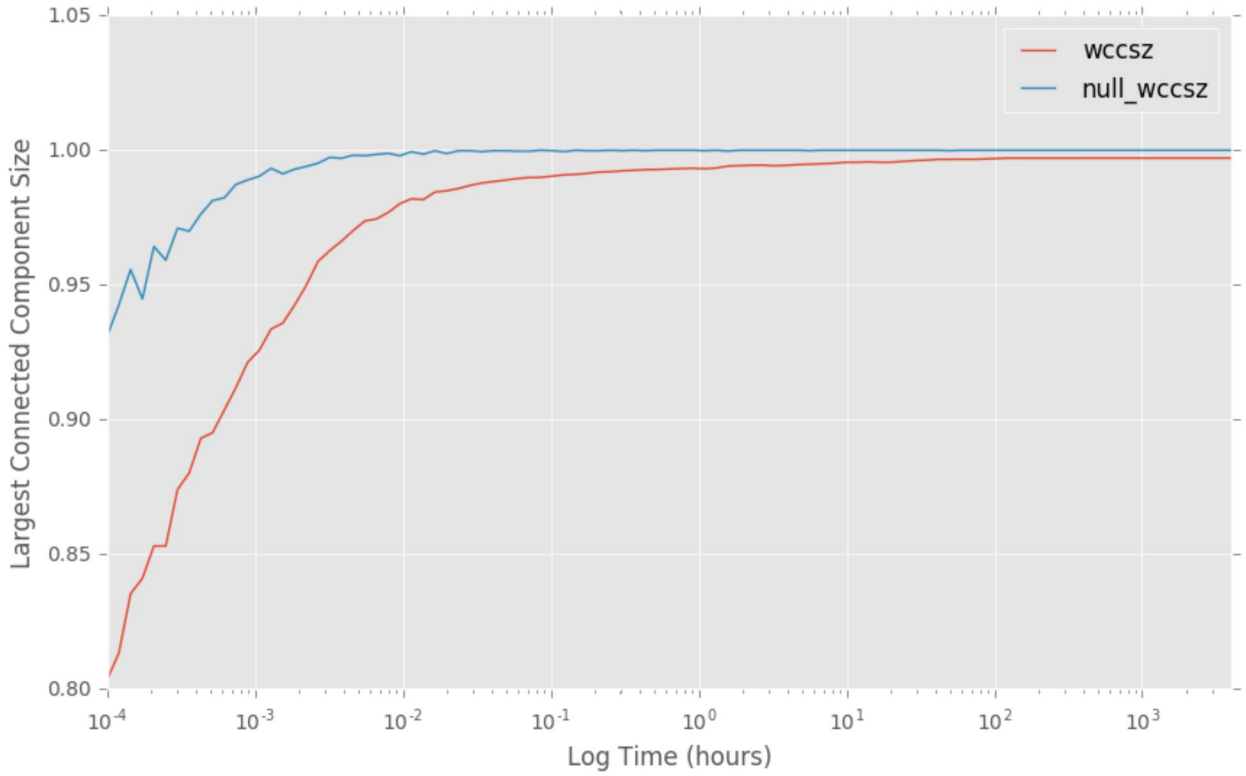


Figure 10. Colors represent size of cumulative time window. Black dots are from the shortest time window (0 second to 0.36 seconds) and yellow dots are from the largest time window (0 seconds to 6 months). Note that the edge weight distribution evolves into a Pareto distribution as the size of the cumulative time window increases.

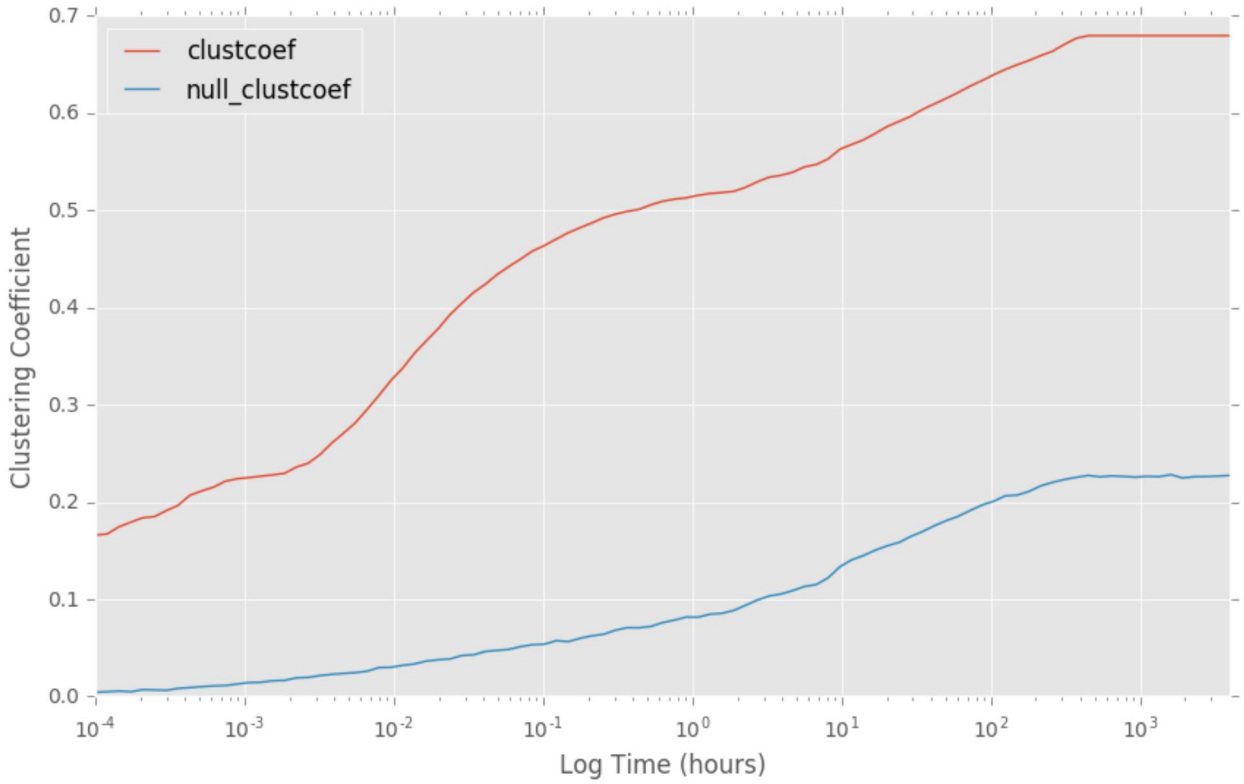
Largest Connected Component and Global Clustering Coefficient

The size of the largest connected component gives us a measure of the connectedness of the graph. We calculated this for each time window and compared it to the corresponding null model.



We see something interesting: the size of the largest connected component in the Sci-Hub network is lower than the null model. This is due to the fact that the Sci-Hub network is extremely dense but broken into clusters. When the edges of the Sci-Hub network are randomized, the clusters are joined together and the size of the largest connected component increases. This is further evidence that the Sci-Hub network is modularized; an idea we explore here and in the next section.

The global clustering coefficient tells us the degree to which nodes tend to cluster together in a graph. In other words, it tells us the likelihood that for any given node, any two of that node's neighbors will be connected by an edge. We calculated the clustering coefficient for all time windows and compared the results to the corresponding null model.



We see that the clustering coefficient for the Sci-Hub network is markedly higher than the clustering coefficient for the null model. This tells us that the Sci-Hub network is modularized; that is, there are multiple densely connected groups of nodes with weak ties to each other. We explore this further in the next section.

Community Structure and Purity

Initially, we calculated graph modularity on a sample of 500 Sci-Hub users to gain insight into the structure of the co-download network. We found that even with such a small sample, the co-download network was highly structured. As expected, the smaller the time window within which pairs of journals were downloaded, the more closely related the journals were. We created a visualization of the Sci-Hub network with nodes and edges colored by community. We used Gephi to visualize the graph. We used the Force Atlas 2 algorithm along with coloration based on graph modularity. All nodes with 1 edge were dropped from this visualization.

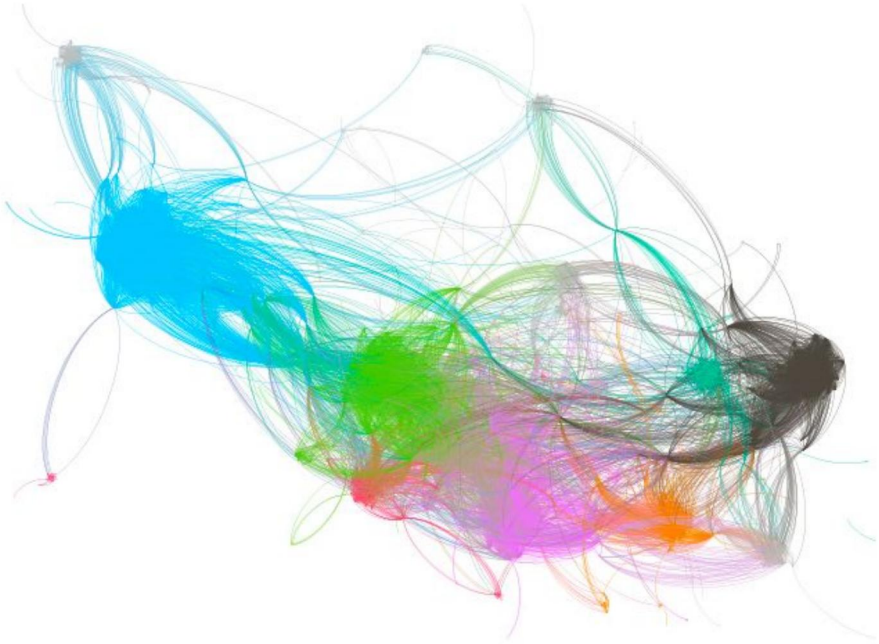
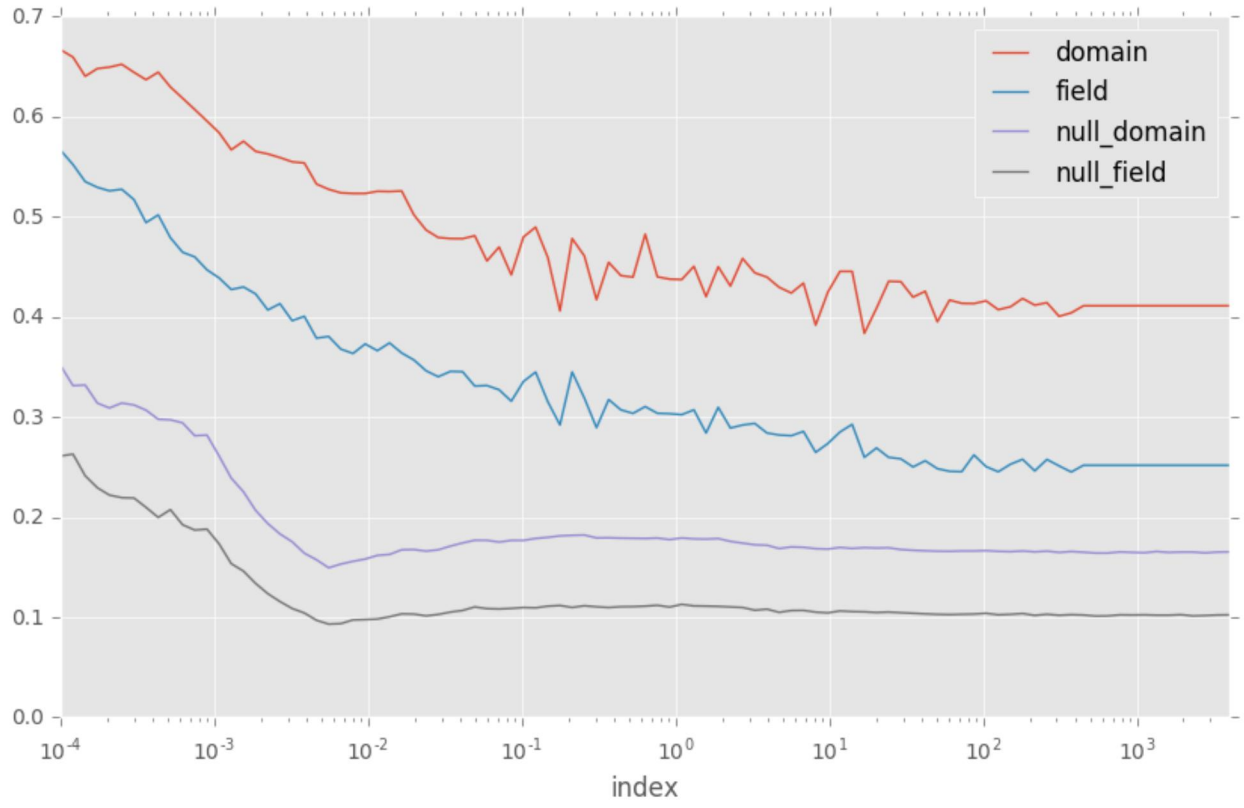


Figure 5. Network community structure on 500 user sample

We built on this analysis by calculating the community structure and purity for each cumulative time window for 1000 users across all time windows. This allowed us to capture the evolution of community structure over time and compare it to a random graph. Note that for all other analyses we used a sample of 10000 users; in this case, due to time constraints it was not possible to perform this analysis with the full sample.

To calculate purity we compare the discovered community structure to a ground truth. The closer the discovered community structure is to the ground truth, the higher the purity. In our case, we used the Science-Metrix database as our ground truth.

In the graph below, we plot community purity at the domain and field level for the Sci-Hub network across all time windows and compare to the purity of the corresponding null model. We do not plot subfield for readability, but the same pattern holds at the sub-field level as well.



Here we have clear evidence in support of our hypothesis. First, we see that the purity of the communities is much higher than the purity of the null models, which shows that there is real community structure at the domain and field level within the Sci-Hub co-download graph. That is, no matter the time window, users are more likely to download papers from similar domains and fields. More importantly, we see a clear decline in purity as the time windows become larger, which is only partially paralleled in the null model. This means that as the co-download time for any two journals becomes larger and larger, the similarity between those two journals at the domain and field level becomes smaller and smaller.

Assortativity

As part of our analysis of how users behave with respect to disciplinary boundaries at expanding time horizons we calculated assortativity for 96 cumulative time-window subgraphs (See Algorithm section). Assortativity measures to what extent edges fall within communities defined by nodes' discrete attributes. The coefficient goes from -1 (where most edges connect nodes from different communities) to 1 (where edges represent homophilous communities). To measure assortativity we calculate $Tr(e) - ||e^2|| / (1 - ||e^2||)$ where e is a category-to-category matrix populated by the proportion of edges that go from one category to another and ||e|| is the sum of all elements in the matrix (Newman 2002: 2). We found that in aggregate users tend to download more within disciplinary boundaries when their time horizons are shorter. This holds for all nested types of boundaries (Domain, Field and Subfield).

A striking result of this analysis is that users will create connections within disciplinary boundaries more readily that published authors at smaller time windows. For instance, it takes about five minutes before users' assortativity at the domain level decays below the authors' measure. Field-level assortativity in Scihub remains above its co-citation counterpart up to the hour-long time window. And Subfield-level assortativity in Scihub is higher than the field-level measure for the co-citation network in the smallest time windows. Moreover, these measures are not an artifact of the number of the number of nodes, degree distribution, or the number of disciplinary categories available in each subgraph, as a comparison to the corresponding null graphs shows.

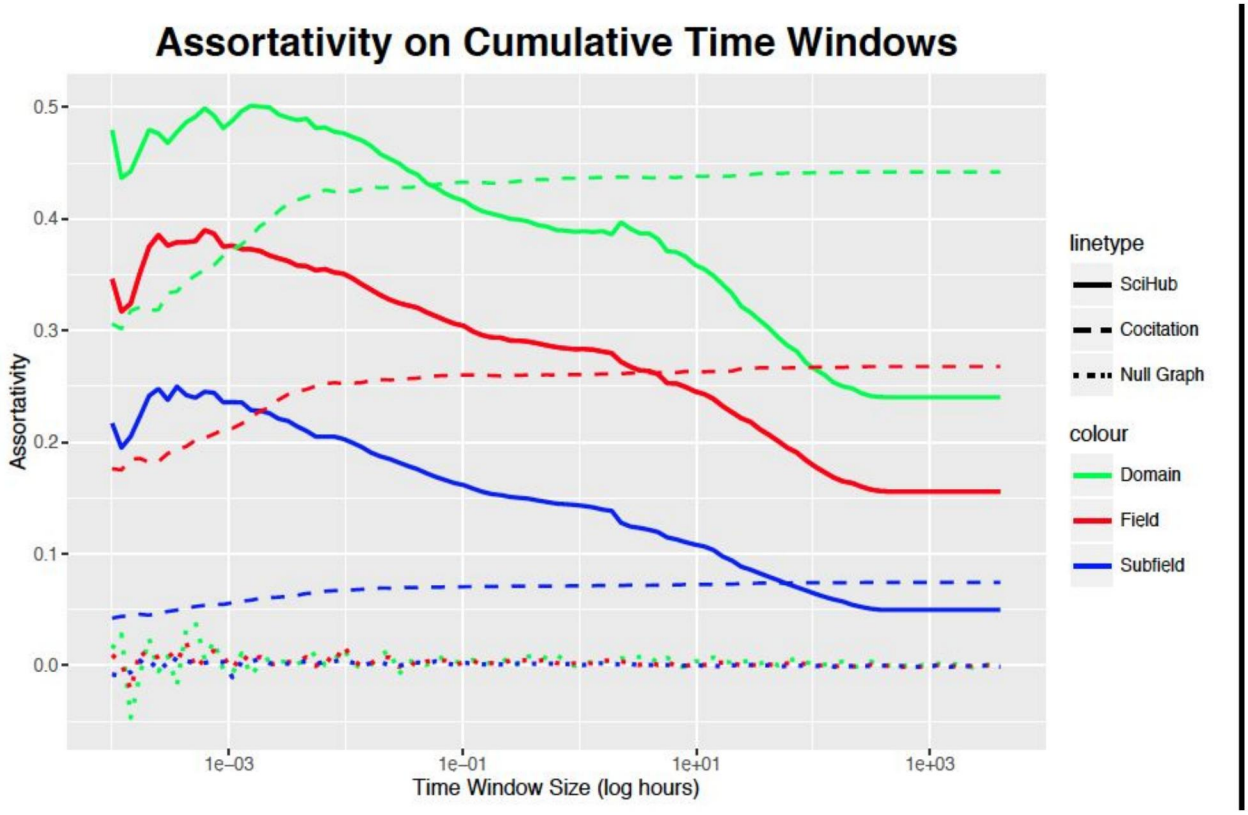


Figure 12. Assortativity across time windows in ten thousand users sample

Validation Measures

A fundamental assumption of our work is that temporal proximity indicates semantic similarity. This means that any given pair of papers or journals can be said to be dealing with similar topics if they have been frequently downloaded together within a short time period. To substantiate this claim we drew a sample of 498714 pairs of co-downloaded papers, then we calculated the temporal proximity between them in log minutes as well as their linguistic proximity in TF-IDF cosine similarity. The latter measure (i) takes the titles of each paper, (ii) counts the frequency of each word, (iii) weights the word counts by the frequency of each term in the entire sample and then (iv) compares the resulting article term-frequency vectors. When comparing cosine similarity and temporal proximity we would expect to find a negative correlation if our assumption is supported by the data. This means that a larger temporal distance should be negatively associated with linguistic proximity. We found that this relation holds in our sample with a -.24 pearson product moment correlation at the paper-paper co-download level.Similarly, we get a negative -0.12 When we aggregate to journal-journal level and train the cosine similarity score on grouped titles for each journal.

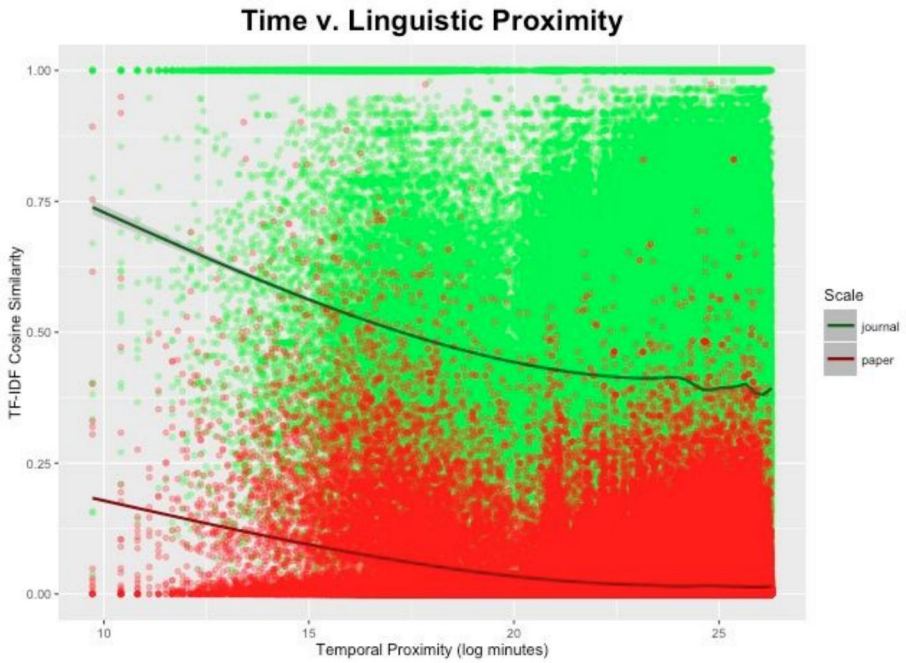


Figure 13. Relation of temporal proximity and linguistic similarity

Discussion

Time is a scarce resource, and not all forms of disciplinary boundary crossing may be possible when time is constrained. Studies of information seeking practices in everyday life have posited that when time is pressing people tend to draw from familiar and convenient information sources (Chatman 1998, Savolainen 2008, Cannaway et al. 2011). Users may engage in a download spell of a few minutes to satisfy an immediate information need at hand, or they may engage in months-long research projects involving hundreds of downloads. The time horizon involved in the more immediate and urgent tasks may require the user to follow clear disciplinary boundaries, while longer projects may present opportunities to cross them. This is supported by the fact that consuming information across boundaries requires an investment in adopting different conceptual frameworks and expert vocabularies (cf. Cultural holes in scientific communication Vilhena et al. 2014).

Our analysis of assortativity supports our claim that given the expansion of time horizons users will forego the convenience and familiarity related to disciplinary boundaries. Moreover, within the shortest of time horizons they will remain more strictly within boundaries than their published counterpart in the co-citation network. Our analysis of community structure provided further evidence for our hypotheses. First, we found that the purity of the communities in the Sci-Hub network is much higher than the purity of the corresponding null models, which shows that there is significant community structure at the domain and field level within the Sci-Hub co-download graph. That is, no matter the time window, users are more likely to download

papers from similar domains and fields. More importantly, we see a clear decline in purity as the time windows become larger. This is partially paralleled in the null model at short time scales, but disappears as time windows become larger. As the minimum co-download time between any two journals becomes larger, the similarity between those two journals at the domain and field level becomes smaller.

To fully understand this result we must take into consideration that Sci-Hub does not have a recommender system and that currently its string-based search is unavailable. This means that to search the Sci-Hub database one must first acquire the DOI of an article. Moreover, to navigate from one paper to the next one must either know beforehand the articles to be consulted, or extract each paper from the bibliography of the last, or pool resources from the recommender systems of official repositories. In any case, Sci-Hub represents a very involved form of information seeking. Therefore there is nothing casual about the disciplinary boundaries followed by Sci-Hub users. In aggregate, these users exhibit a thorough understanding of the topology of science, and the corresponding activity graph is an accurate reflection of that topology.

Next Steps

We intend to streamline our algorithm to work with larger samples of users and to acquire usage data from official repositories for more and better comparisons. More data will also allow us to refine our validation measures. Sampling abstracts and bibliographies in Web of Science and Scopus will be a first step towards more robust measures of the relation between linguistic similarity and temporal proximity.

Assortativity and community detection allowed us to identify the relationship between disciplinary boundaries and users' information seeking behavior. However, another set of instruments would be necessary to identify strategies of boundary-crossing. Statistical models that model changes over time (e.g., Hazard models, hidden Markov models) provide promising ways to analyze sequences of downloads across boundaries.

Bibliography

Archambault**,** É., Beauchesne, O**.,** Caruso J (2011) Towards a multilingual, comprehensive and open scientific journal ontology**.** Proceedings of the 13th international conference of the international society for scientometrics and informetrics 66-77

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

Savolainen, R. (2006). Time as a context of information seeking. *Library and Information Science Research*, *28*(1), 110–127.

Savolainen, R. (2008). Everyday information practices. Lanham, MD: Scarecrow

Connaway, L.S. Dickey, T., Radford, M. (2011) "If it is too inconvenient I'm not going after it:" Convenience as a critical factor in information-seeking behaviors Library & Information Science Research 33 (2011) 179–190

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

Vilhena, D. A., Foster, J. G., Rosvall, M., West, J. D., Evans J., Bergstrom C. T. Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication

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

Whiting, S. W. (2015) Temporal dynamics in information retrieval. PhD thesis. Glasgow: University of Galsgow http://theses.gla.ac.uk/6850/

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

Figures Annex