Academic Torrents: A Community-Maintained Distributed Repository

Joseph Paul Cohen
The University of Massachusetts Boston
100 Morrissey Blvd. Boston, MA
joecohen@cs.umb.edu

Henry Z Lo
The University of Massachusetts Boston
100 Morrissey Blvd. Boston, MA
henryzlo@cs.umb.edu

ABSTRACT

Fostering the free and open sharing of scientific knowledge between the scientific community and general public is the goal of Academic Torrents. At its core it is a distributed network for efficient content dissemination, connecting scientists, academic journals, readers, research groups, and many others. Leveraging the power of its peer-to-peer architecture, Academic Torrents makes science more accessible through two initiatives. The open data initiative allows researchers to share their datasets at high speeds with low bandwidth costs through the peer-to-peer network. The cooperative nature of scientific research demands access to data, but researchers face significant hurdles making their data available. The technical benefits of the Academic Torrents network allows researchers to scalably and globally distribute content, leading to its adoption by labs all around the world to disseminate and share scientific data. Academic Torrent's open access initiative uses the same technology to share open access papers between institutions and individuals. We design a connector to our network that acts as a onsite digital stack to complement the already existing physical stack curated in the same manner. Utilizing the collective resources of the academic community we eliminate the biases in the closed subscription model and the pay to publish model.

Categories and Subject Descriptors

E.5 [**Data**]: Files—*Organization structure*; C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design; H.2.7 [**Database Management**]: Database Administration—*Data warehouse and repository*

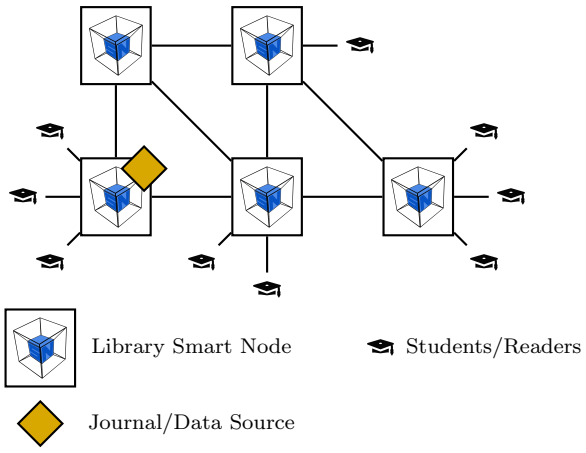

Figure 1: An example network of curated Library Smart Nodes which provides the dissemination infrastructure for a journal or similar data source.

1. INTRODUCTION

Scientific knowledge is largely publicly funded, but only available to a select group of individuals. This exclusiveness hurts the legitimacy and progress of the scientific community, as argued by many leading researchers [3, 2]. The problem stems from both closed access publishing and data sets being made unavailable, both of which hurt the flow of information. The aim of Academic Torrents (AT) is to help solve these problems by making sharing easier and more scalable.

Though barriers are different for sharing papers and datasets, AT addresses both sets of problems using one common architecture.

- Sharing large datasets is limited by storage space requirements, speed, and availability.
- Sharing papers is limited by systematic problems in academic publishing coupled with the cost burden of publishing.

We present a framework for scalably and globally distributing content. The network nodes adapt to maximize content availability, while taking into account each node's interest in only hosting particular types of data. By using multiple levels of content curation, we allow easy selection of desired content and efficient exclusion of undesired content.

	Maintenance	Limited Bandwidth	Speed	Robust	Cost	Complexity
Single Server	Moderate	Yes	Moderate	No	Moderate	Low
Apache Style Mirroring	High	Moderate	Moderate	Moderate	Moderate	High
Mailing Media	High	Yes	Slow	No	Low	Low
Free Repositories	Low	Yes	Slow	Moderate	Free	Low
Proprietary Repositories	Low	Moderate	Moderate	Moderate	High	Low
Academic Torrents	Low	No	High	Yes	Low	Moderate

Table 1: We compare current data dissemination systems. The only weakness of Academic Torrents is complexity; current effort is focused on making the platform easier to use.

2. SHARING DATA

Though existing solutions, as seen in Table 1, can make sharing data easier for researchers, one common problem is that they are hard to scale. With more users, existing solutions lead to speed and availability problems. Conversely, using the BitTorrent protocol, an increasing user base actually improves network performance [1], with an added bonus of file availability. We use this peer-to-peer protocol as the basis of Academic Torrents (as did BioTorrents [4]).

Academic Torrents (AT) provides a searchable index for content. Users can organize content relevant to them into collections, which allows them to locate or mirror desired content. "Direct Numerical Simulation of Turbulent Flows" [1] is one such collection. This page contains a description and list of all entries relating to this topic, and is curated by a researcher in this area.

We also provide an API for integrating AT with other dissemination systems, so that AT can enhance existing distribution methods by adding a peer-to-peer pathway. We have created tools that use this API in order to quickly upload data via scripts or a graphical interface. In order to aid in migration and integration, we utilize the BitTorrent "Web Seed" specification to allow existing HTTP or FTP servers to use Academic Torrents directly without modification.

The soundness of this idea has led many individuals and institutions to offer resources for mirroring data on the AT network. This has raised a question - how best to allocate these donor resources to balance supply and demand? For example, some donors can only donate so much space - how should the donor fill up this space to best ensure dataset availability? To investigate questions like these, the AT team is currently developing the *Smart Node*. This software will run on donor computers and manage resources, e.g. disk space and bandwidth, in an intelligent and dynamic way to both maximize file availability and download speed globally on the AT network. Donors curate and select collections of content they want to mirror on the network. The goal of the Smart Node is to make the Academic Torrents network more robust by balancing dynamically network users' needs and the resources they have.

3. DISTRIBUTED PUBLISHING

Closed subscription publishing has restricted information from the general population and even many universities. To counter this, a large Open-Access model has been championed, but because authors are asked to pay it introduces a new bias which even more unacceptable.

Some journals such as NIPS, JMLR, and PVLDB publish and disseminate their papers without charge. A major challenge with disseminating research publications is the burden of ongoing management and hosting. AT is designed to share the burden of disseminating research.

To contrast Closed Access publishing and Open Access publishing we propose Distributed Publishing. This model allows journals to share the burden of overhead that is incurred by having their readers share in hosting the content. This lowers the required uptime of the journal and, unlike fees or donations, allows the costs of journal infrastructure to be under the direct control of the reading institution.

Libraries currently contain librarian curated physical stacks to store books onsite. We model our system as a librarian curated "digital stack" which we implement with a Library Smart Node (LSN) that fulfills two roles:

- Librarians select collections, such as journals or theses, to mirror on their LSN. This allows them to maintain copies of these documents, participate in dissemination of them, and provide a rich document viewing experience for their community.
- Journal editors index their publications on AT in an automatic fashion allowing others to mirror and assist in content delivery.

The example distributed publishing network shown in Figure 1 allows users to access the data source without directly contacting it. Each reader can automatically verify their content is authentic even though it does not come from the source directly. Also, any node can be under maintenance and the content can still be accessed.

4. CONCLUSION

Academic Torrents is using peer to peer technology to solve many of the problems in the academic community: creating tools for a globally distributed storage network with Smart Nodes, creating user-friendly software to interact with the network, and using this infrastructure to tackle the problems in academic publishing.

5. ACKNOWLEDGMENTS

This work is partially funded by The National Science Foundation Graduate Research Fellowship Program (Grant No. DGE-1356104).

6. REFERENCES

[1] D. Erman, D. Ilie, and A. Popescu. Measuring and modeling the BitTorrent content distribution system. *Computer Communications*, 33, Supplement 1:S22–S29, Nov. 2010.

[2] M. Jordan. Leading ML researchers issue statement of support for JMLR, Oct. 2001.

[3] D. E. Knuth. Donald knuth: Editorial board, journal of algorithms, Oct. 2003.

[4] M. G. I. Langille and J. A. Eisen. BioTorrents: a file sharing service for scientific data. *PLoS ONE*, 5(4):e10071, Apr. 2010.

[1] http://academictorrents.com/collection/direct-numerical-simulation-of-turbulent-flows