DATA 620: Final Proposal

Building a pipeline for exploring and visualizing authorship and publishing practices in astrophysics

Daina Bouquin, John DeBlase

Background on data sources

The SAO/NASA Astrophysics Data System (ADS), is an online database of over eight million astronomy and physics papers from both peer reviewed and non-peer reviewed sources. The ADS is a highly used resource in the Astronomy and Physics communities and has many levels of indexing. The ADS API makes it possible to query this valuable resource to better understand authorship and publishing behavior in these fields among many other applications. In the near future, the ADS plans to begin incorporating Unified Astronomy Thesaurus (UAT) keywords into their indexing schema. The ADS is managed by the Smithsonian Astrophysical Observatory at the Harvard–Smithsonian Center for Astrophysics.

Problem formulation

Are there notable differences in centrality measures among networks of authors and publications in various astrophysical domains?

Plan

Our goal is to build a prototype web framework to automate a data pipeline for building large graphs and interactive visualizations showing authorship and publication tendencies in astronomy and astrophysics. This tool will be used to study a specific situation related to the ADS, however the design of the tool could potentially be applied to any graph related problem.

Specifically, we will create a python app using the ADS API, Flask, SQLite, Ajax, and sigma.js using the following logic:

Data Warehousing:
ProccessADS(script) → inspect → ProcessGraph(script) → inspect/export to csvs → update_schema/load_tables SQLite on server (script)

Clientside:
Form data (query strings/fields from dropdowns etc) → Ajax request → Python processing endpoint (Flask route function)

Inside Endpoint(from AJAX request):
Pull from database → Prepare object for sigma/client → parse into JSON response →jsonify(object) to client

Clientside:
Receive JSON object → sigma.js processing and display → any additional functionality

Fields to be compared (top level terms from the UAT)

- Astrophysical processes
- Cosmology
- Exoplanet astronomy
- Galactic astronomy
- Extragalactic astronomy
- High energy astrophysics
- Interstellar medium
- Interdisciplinary astronomy
- Observational astronomy
- Solar astronomy
- Solar system astronomy
- Stellar astronomy

Roles

Daina - proposal, domain-specific decisions, network analytics scripting, write-up
John - converting scripts into python app/flask protocol, sql database, sigma.js
Both - testing, iterating

Limitations

ADS has not yet implemented UAT indexing. Therefore the initial API queries will result in “full text” searches as opposed to results from comprehensive topical indexing. Once UAT keywords are implemented the initial query can be amended. Additionally, because the analysis will be process-intensive, data warehousing will be executed resulting in a “snapshot” visualization from the time that the query was run. The visualization will not be based on a “live” query of the ADS.