



**Smithsonian/NASA
Astrophysics Data System**

The Search for Structure in the Astronomical Bibliosphere

Alberto Accomazzi

ADS

Harvard-Smithsonian Center for Astrophysics

aaccomazzi@cfa.harvard.edu



Smithsonian/NASA Astrophysics Data System

Overview

- ADS Bibliographic Data and Use
- Uses and applications of networks on bibliographies
- Methodology
- Current results and prototype
- Blue Sky Stuff
- Conclusions



Smithsonian/NASA Astrophysics Data System

ADS Bibliographic Data

- Bibliographic records of scholarly articles in 3 databases:
 - Astronomy: 1.5M (700K w/abstracts, 500K fulltext)
 - Physics: 3.9M (2.6M w/abstracts)
 - ArXiv e-prints: 500K (all w/abstracts)
- Bibliographic Groups:
 - Major missions, institutions
 - Links to datasets, Observing proposals
- Over 27M citations
 - Coverage: both Astronomy and Physics



Smithsonian/NASA Astrophysics Data System

ADS Use

- User community:
 - Over 30K regular users (20K registered)
 - 1.2M users from search engines (mostly google)
 - Usage logs since 1996
- Customizations:
 - Login system
 - Private libraries
 - Readership history
 - myADS notification service
 - **example**



Smithsonian/NASA Astrophysics Data System

What are we after?

- Identify semantic structure and flow of ideas in literature
 - Research topics (subject)
 - Communities (co-authorship)
 - Emerging research fields (readership)
 - Connection to datasets
- Create, use and provide tools
 - Bibliographic Classification / Data mining techniques
 - Recommendations / Alerts / Updates
 - Visual interaction tools for end-users



Smithsonian/NASA Astrophysics Data System

How is it done?

- Create a semantic network
 - Define nodes (vertices)
 - Define relationships (edges)
 - Optionally, define weights along edges
- Run a clustering algorithm over the network
 - Directed or un-directed, depending on relationships
 - Lots of options for weights
- Figure out meaning of results
 - Identify and name clusters



Smithsonian/NASA Astrophysics Data System

Network Clustering Algorithm

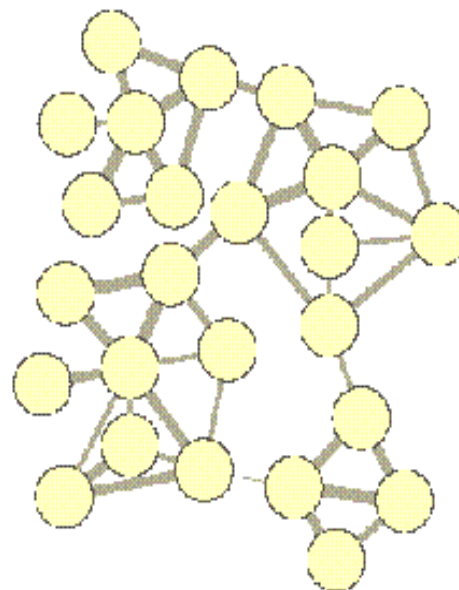
- Rosvall and Bergstrom, arXiv:0707.0609 “Maps of random walks on complex networks reveal community structure”
 - Retain important information, but simplify the network by breaking it into modules
 - Retain and highlight underlying structure by compressing information flow
- Several other algorithms exist, this study is preliminary



Smithsonian/NASA Astrophysics Data System

The Network

- Nodes connected by edges (links)
- Size of edges proportional to information “flow” between nodes
- Flow can be directional

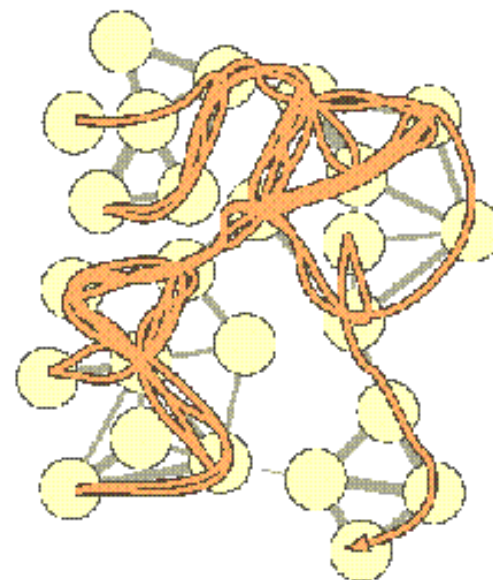




Smithsonian/NASA Astrophysics Data System

Random walk

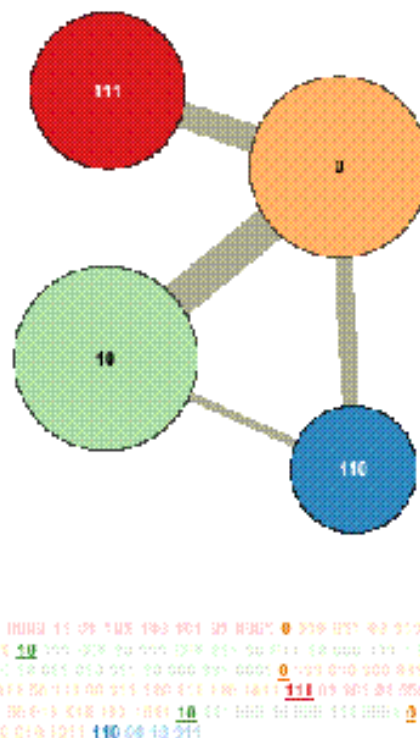
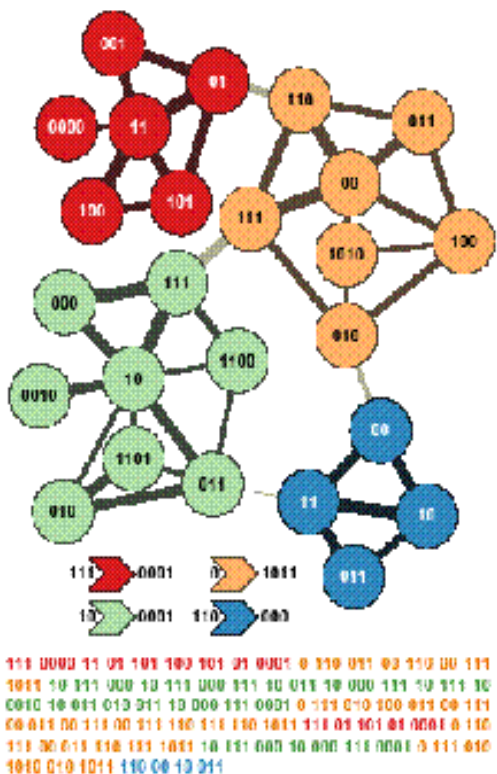
- Sample trajectory of a random walk, as a proxy for information flow in the network
- It uses all the information in the network and nothing more





Smithsonian/NASA Astrophysics Data System

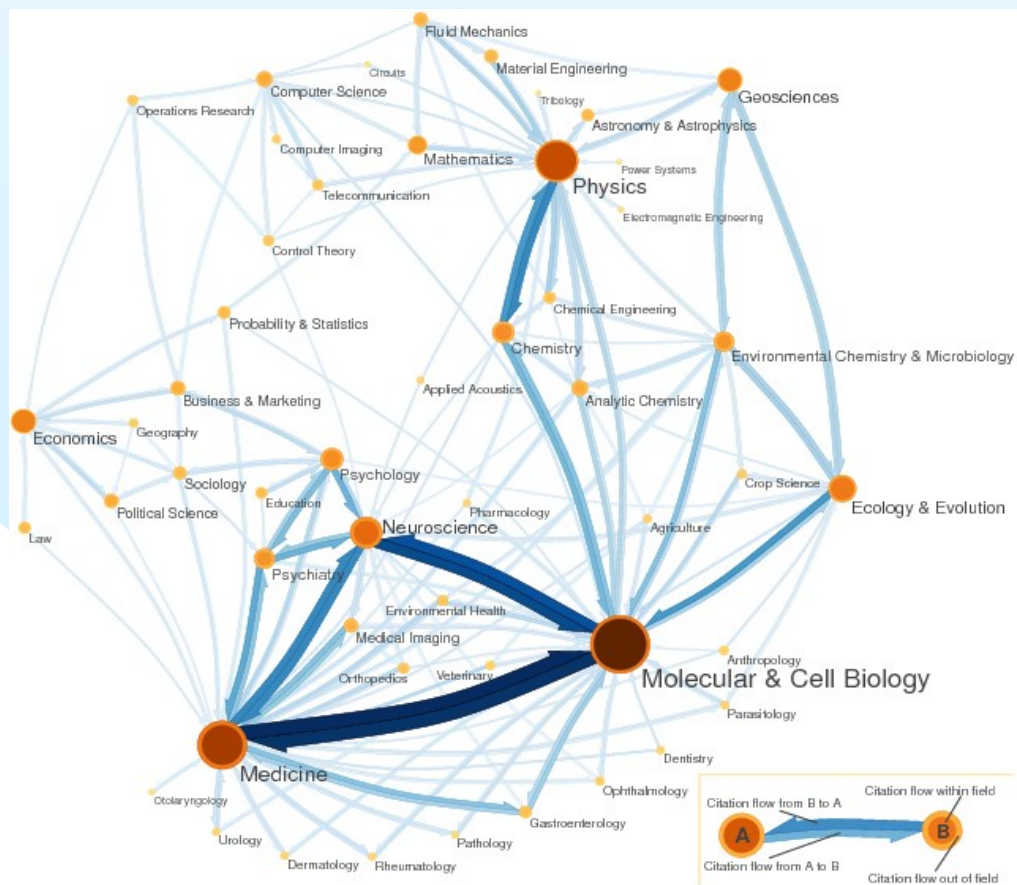
Algorithm (continued)





Smithsonian/NASA Astrophysics Data System

www.eigenfactor.org





Smithsonian/NASA Astrophysics Data System

Application to ADS (1)

- Network of papers / citations:
 - Articles in major astronomical articles published in 2000-2005
 - Nodes: papers (35,941)
 - Edges: citations (325,353)
 - Clustering yields over 20,000 sets
 - **Map of top clusters**
 - **Interactive Applet**



Smithsonian/NASA Astrophysics Data System

Application to ADS (2)

- Network of papers / subject headings
 - Articles in major astronomical articles published in 2000
 - Nodes: papers (14,810)
 - Edges: keyword co-occurrences (7,325,490)
 - Clustering yields 38 modules
 - **Map of connected sets**
 - **Interactive applet**



Smithsonian/NASA Astrophysics Data System

Next steps

- Further research in clustering algorithm needed
- Robustness and adaptability need to be addressed
- Classification of new papers (nodes) in network and cluster set
- Browsing tool should allow both coarse and fine-grained view of cluster space
- View and analysis of a subset of data within the network (“where do I fit in the map?”)



Smithsonian/NASA Astrophysics Data System

Other Possible Networks

- Papers based on word co-occurrence
- Papers based on co-readership
- Papers based on co-observation
- Papers based on co-authorship
- Datasets based on co-occurrence in publications
- Objects based on co-occurrence in publications
- Higher order relationships based on publication web



Smithsonian/NASA Astrophysics Data System

What other semantic stuff can/could ADS do?

- Create richer user profiles based on user “bookmarks” of publications (private libraries)
- Provide suggestions based on aggregate user readership (“Also-read” papers).
- Help uniquely identify entities (e.g. Authors, projects, institutions)
- Bridge the missing links between data products by virtue of their links to publications (e.g. Observing proposals)



Smithsonian/NASA Astrophysics Data System

Want to participate?

- Provide expertise in network and cluster analysis
- Collaborate on user-interfaces for network navigation
- NLP techniques for data mining and keyword extraction
- Author / Affiliation disambiguation
- Topic analysis and research trends
- Creation and maintenance of “semantic” links between bibliographies and data / objects
- Integration of IVOA SSO/OpenID with ADS user database



Smithsonian/NASA Astrophysics Data System

Thank you