Scholarly Big Data Information Extraction and Integration in the CiteSeerX Digital Library

Kyle Williams, Jian Wu, Sagnik Ray Choudhury, Madian Khabsa and C. Lee Giles

Information Sciences and Technology
Computer Science and Engineering
The Pennsylvania State University

CiteSeer$^{x}_{\beta}$

PENNSTATE
1855

Scholarly Big Data

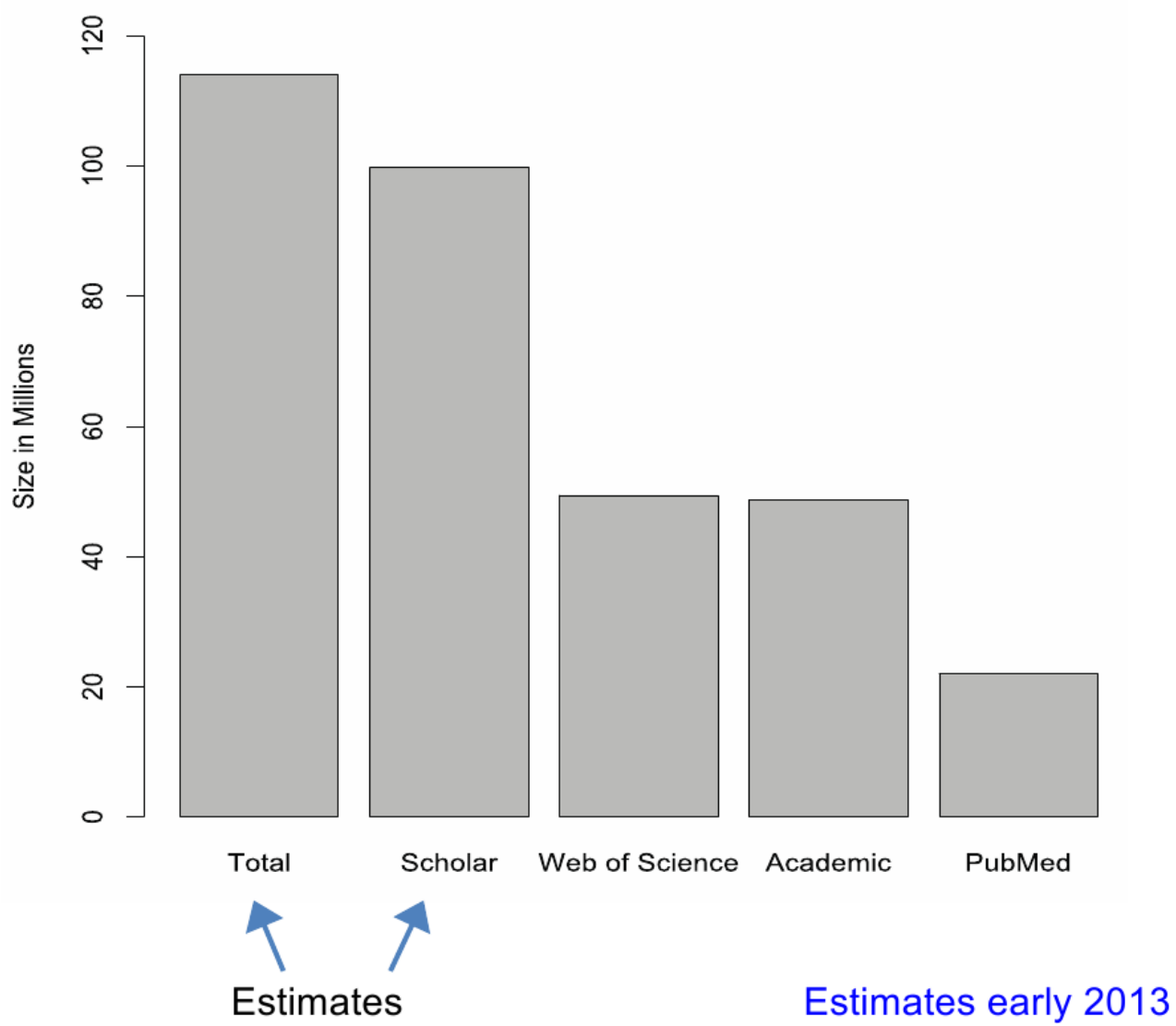
- Scholarly Big Data includes all academic research output
  - Journal and conference publications
  - Books
  - Theses
  - Slides, data, course materials…

- Often found in data repositories
  - Google Scholar, Arxiv, Microsoft Academic, CiteSeerX, PubMed, University Libraries, etc.

Scholarly Big Data

- Scholarly Big Data includes all academic research output
  - Journal and conference publications
  - Books
  - Theses
  - Slides, data, course materials…

- Often found in data repositories
  - Google Scholar, Arxiv, Microsoft Academic, CiteSeerX, PubMed, University Libraries, etc.



Estimate of amount of scholarly documents on the Web

How Much is Available?

- Estimate at least 114 million documents
  - Google Scholar has nearly 100 million
- At least 24% of these documents are publicly available
  - 27 Million
  - Though it varies a lot by field
- Estimates that 43% of articles published between 2008-2011 are freely available online (Archambaul et al., 2013)

PENNSTATE
1855

Why is it Useful?

- Analyze scholarly and research trends
- Evaluation of investments in science and scholarship
- Identify opportunities for collaboration
- Evaluate individual scientist, groups and organizations

PENNSTATE
1855

Challenges

- Integration from multiple data sources
- Heterogenous data
- Duplication
- Entity Linking and disambiguation
- Scalability
- Reuse

The CiteSeerX Digital Library

- A (almost) fully automated digital library and search engine for scholarly big data
  - Focused crawling to retrieve and integrate scholarly data
- Information Extraction
  - Metadata, citations, figures, tables, acknowledgements, algorithms
- Entity linking and disambiguation
  - Document and citation clustering
  - Name disambiguation

CiteSeer$^{x}_{\beta}$

PENNSTATE
1855

TABLE I

COLLECTION AND USAGE STATISTICS FOR CITESEER$^{\chi}$

Statistic	**Value**
#Documents	3.5 million
#Unique documents	2.5 million
#Citations	80 million
#Authors	3-6 million
#docs added monthly	300,000
#docs downloaded monthly	300,000-2.5 million
Individual Users	800,000
Hits per day	2-4 million

Focused Crawling

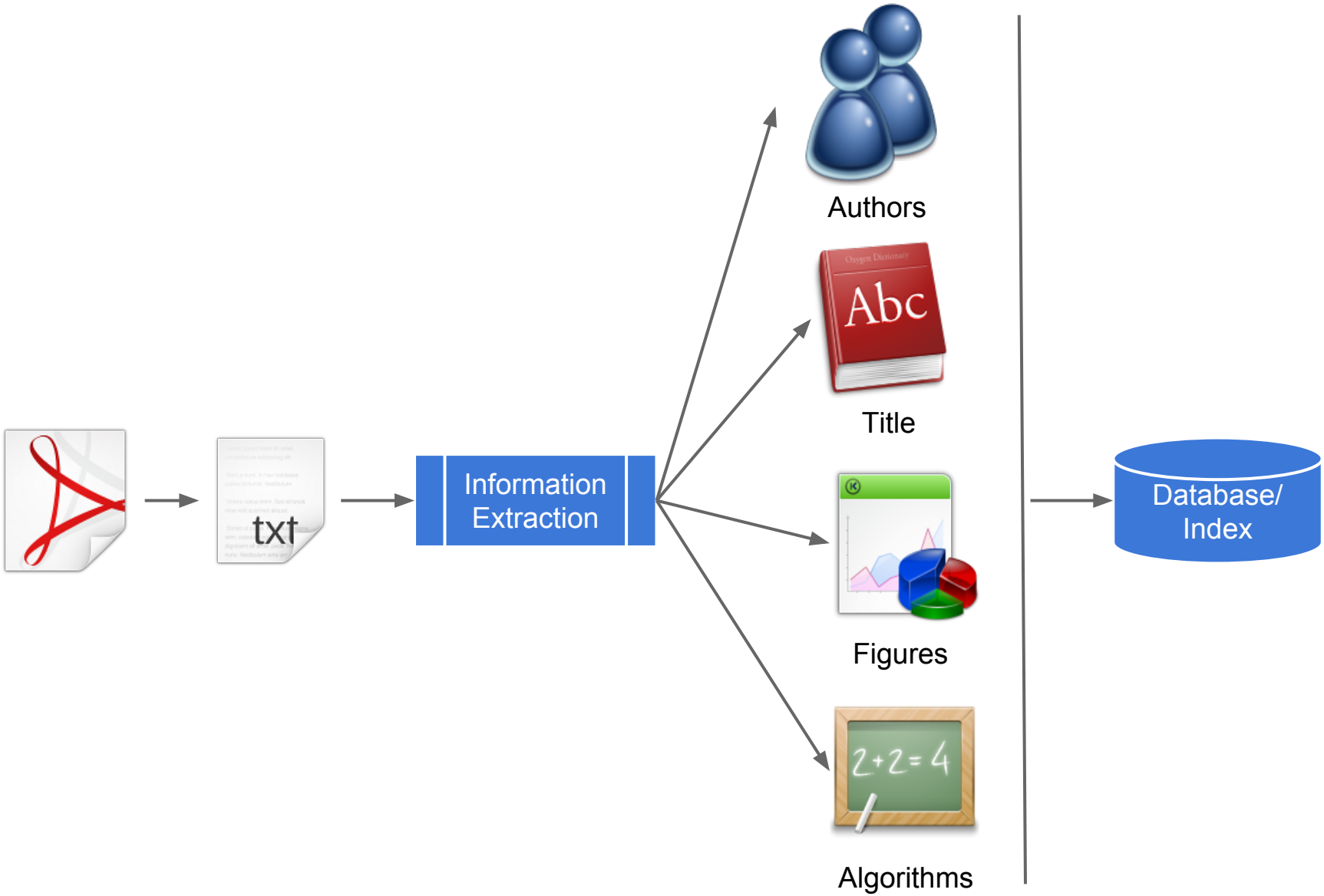
- CiteSeerX crawls freely accessible scholarly data on the Web
  - Maintain a whitelist of URLs that are okay to visit and are likely to be good links
  - Only retrieve PDF and PostScript files
- 50,000-100,000 document retrieved per day
  - Document classification module based on regular expressions to only keep academic documents
- Aggregates and integrate all documents in a single location

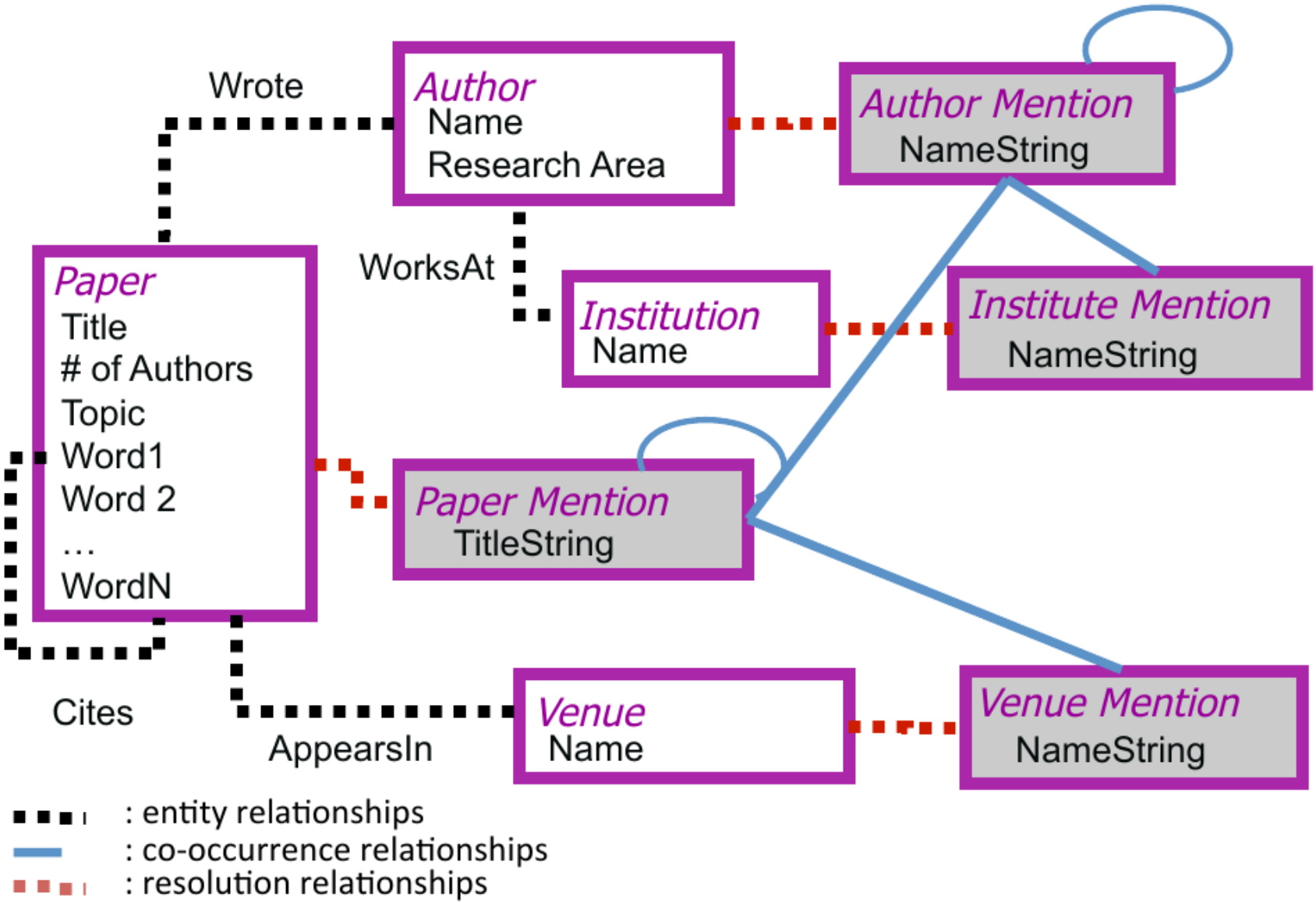
PENNSTATE
1855

Information Extraction

- Fully automated information extraction from scholarly documents
  - Different types of information that can be extracted
  - Used for search, browsing, services, research
- All extracted data is indexed and integrated for search via CiteSeerX
- Needs to be robust to variations among scholarly documents

PENNSTATE
1855


txt
Information Extraction
Authors
Abc
Title
Figures
2+2=4
Algorithms
Database/ Index



Source: Getoor & Machanavajjhala, KDD'13

Header Extraction

- Contains some of the most important information in scholarly documents
  - Authors, title, abstract, venue, etc.
- Performed using SVM-based extractor
  - Tags each entity
- Sometimes errors or missing information
  - Allow for user corrections on CiteSeerX website
  - Match and link records with DBLP

PENNSTATE
1855

Citation Extraction

- One of the key features of CiteSeerX
- Allows for autonomous citation indexing
  - Build up citation statistics for authors, venues, etc
  - Calculate h-index
- Citation extracted using CRF-based tool
  - Each entity in a citation string is tagged

PENNSTATE
1855

Other IE

- Scholarly data has many other sources of information
  - CiteSeerX extracts figures, acknowledgements, tables and algorithms from papers
- Build services based on this extracted data
  - AckSeer, TableSeer, AlgorithmSeer, etc
- All data is all linked to the original source

PENNSTATE
1855

Challenges for Automatic IE

- Accuracy of automatic IE
  - Needs to work on heterogeneous input data
  - Can be improved through linking
  - Coverage/accuracy
- Scholarly data is growing and scalability poses a major challenge
  - Improved algorithms?
  - Distributed extraction

PENNSTATE
1855

Deduplication and Clustering

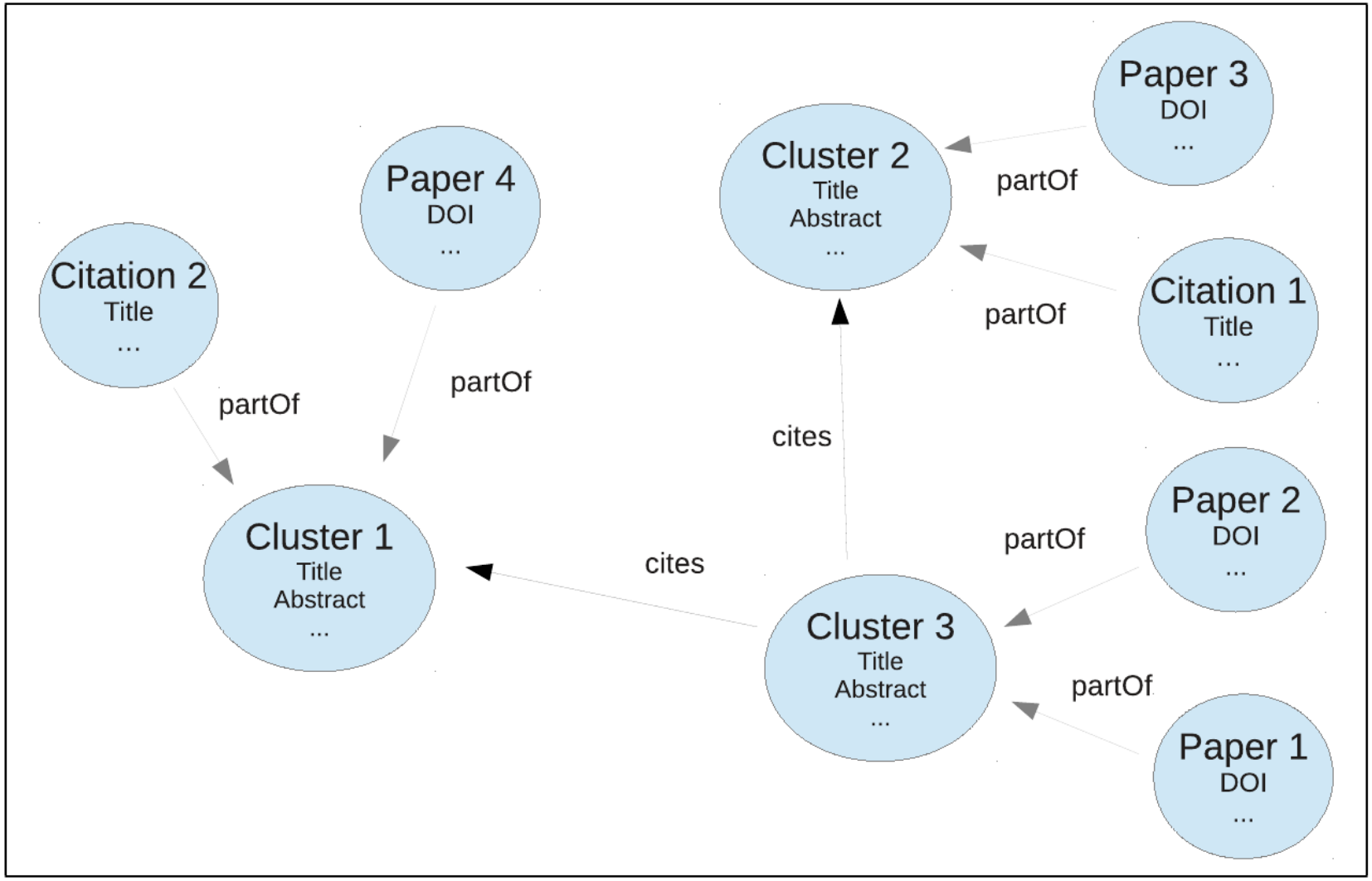
- Multiple versions of papers might exist on the Web that have minor differences
  - Estimate that as many as 30% of papers in CiteSeerX might be duplicates
  - Not bitwise identical
- Cluster near duplicates under a single ID based on automatically extracted metadata
  - But this can be wrong with bad metadata extraction
  - Looking into other more advanced methods

PENNSTATE

Citation Clustering and Linking

- Like duplicate papers, many citations exists to single papers
- Match citations to papers
  - Allows us to calculate citation statistics
  - Assign to same clusters as papers
- Citations can be used to improve paper metadata

PENNSTATE
1855



PENNSTATE
1855

Author Name Disambiguation

- Classic problem in digital libraries

  “C. L. Giles” vs “Lee Giles”

- Important to allow proper attribution
  - Citations, contributions, search, etc.

- Build a similarity profile for each author and cluster authors

PENNSTATE
1855

Sharing Data

- Important to foster collaboration and research

- Sharing data is challenging
  - Copyright issues
  - Take down notices
  - Big data - 6TB and growing by 10-20GB per day
  - How to share and distribute this?

- Make metadata available via OAI-PMH
  - About 5000 requests per month
  - Also make data available on Amazon S3

PENNSTATE 1855

Sharing Code

- We open source all our software
  - Allows other groups to run their own versions of CiteSeerX and make improvements
- CiteSeerX forked 8 times since moving to GitHub

https://github.com/SeerLabs

PENNSTATE
1855

Services

- Moving towards a service platform
- People may not be interested in running CiteSeerX, but rather integrating CiteSeerX components into their own workflows
- CiteSeerExtractor provides API access to information extraction modules
- Repository API provides access to repository

PENNSTATE
1855

Conclusions

- Scholarly big data is rich and useful
  - Linking entities provides an opportunity to better understand scholarly undertaking
- But integrating scholarly data is challenging due to heterogeneity and scale
- Research opportunities in information extraction, integration and entity linking

PENNSTATE
1855

Thanks

- CiteSeerX and this work is supported by the National Science Foundation.

- The many contributors to CiteSeerX

PENNSTATE
1855