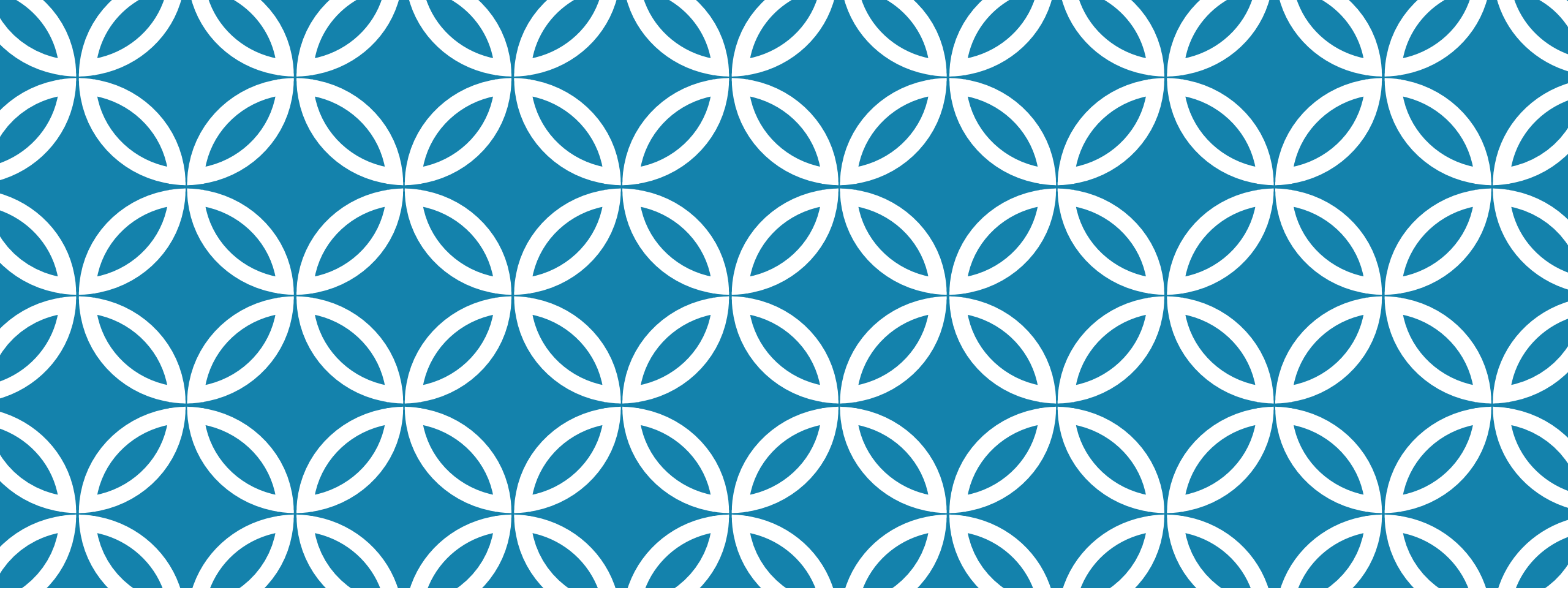


LECTURE 5

Win+w

Signal Processing on Databases

Jeremy Kepner

Lecture 1: Using Associative Arrays

This work is sponsored by the Department of the Air Force under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

Outline

- Citation Data
  - Schema
  - Pipeline
  - Observations
- Graph Construction
- Multi-Hyper Graphs
- Summary

Exploded Schema (Key Table)

Input Data

ut	auth	docid	ref.docid
1234	a		a
1243	b	b	
4321		c	c

Accumulo Table: TkeyT

	ut/1234	ut/1243	ut/4321
auth/a	1		
auth/b		1	
docid/b		1	
docid/c			1
ref.docid/a	1		
ref.docid/c			1

	auth/a	auth/b	docid/b	docid/c	ref.docid/a	ref.docid/c
ut/1234	1				1	
ut/1243		1	1			
ut/4321				1		1

Accumulo Table: Tkey

- Holds structured citation data
- Primary table for constructing graphs
- Values hold position in record (i.e. 1st, 2nd, 3rd author/reference …)

Exploded Schema (Txt Table)

Accumulo Table: Ttxt

	ref	title	abstract
ref.docid/ 1234	a		
ut/1243	b	b	b
ut/4321		c	c

- Traditional table for holding long formatted reference, title, and abstract strings
- Eliminates inconvenient long strings from key table
- Typically only used for manual verification

Exploded Schema (Ngram Table)

Input Data

ut	title	abstract
1234	a b a ...	c d ...
1243	b ...	
4321		d ...

Accumulo Table: TngramT

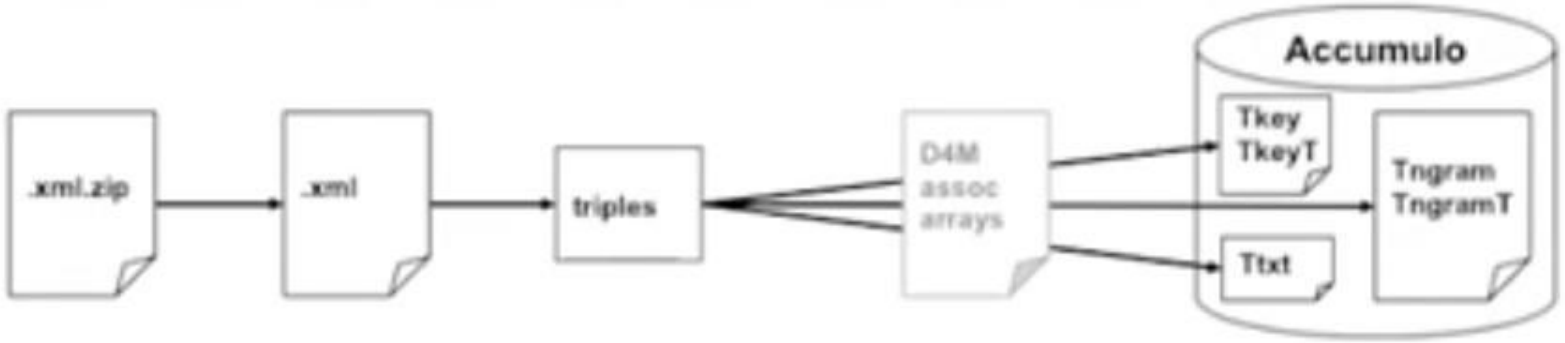
	ut/1234	ut/1243	ut/4321
title/1gram/a	1,3	1	
title/1gram/b	2		
title/2gram/a b	1		
abstract/1gram/c	1		
abstract/1gram/d	2		1
abstract/2gram/c d	1		

Accumulo Table: Tngram

	title/1gram/a	title/1gram/b	title/2gram/a b	abstract/1gram/c	abstract/1gram/d	abstract/2gram/c d
ut/1234	1,3	2	1	1	2	1
ut/1243	1					
ut/4321					1	

- Holds 1, 2, 3-grams for titles and abstract (5x larger than key table)
- Values hold word position(s) in record
- Separation allows ngram ingest to be done independently

Typical Processing Chain



1. Uncompressed XML file [once]
2. Read XML into binary structure and parse into triples [a few times to finalize parse code]
3. Construct D4M associative arrays from triples to check data [once]
4. Insert triples into Accumulo [once per database]

- Used several intermediate files so that fewest steps need to be redone during development

Single Node 42M Record Times

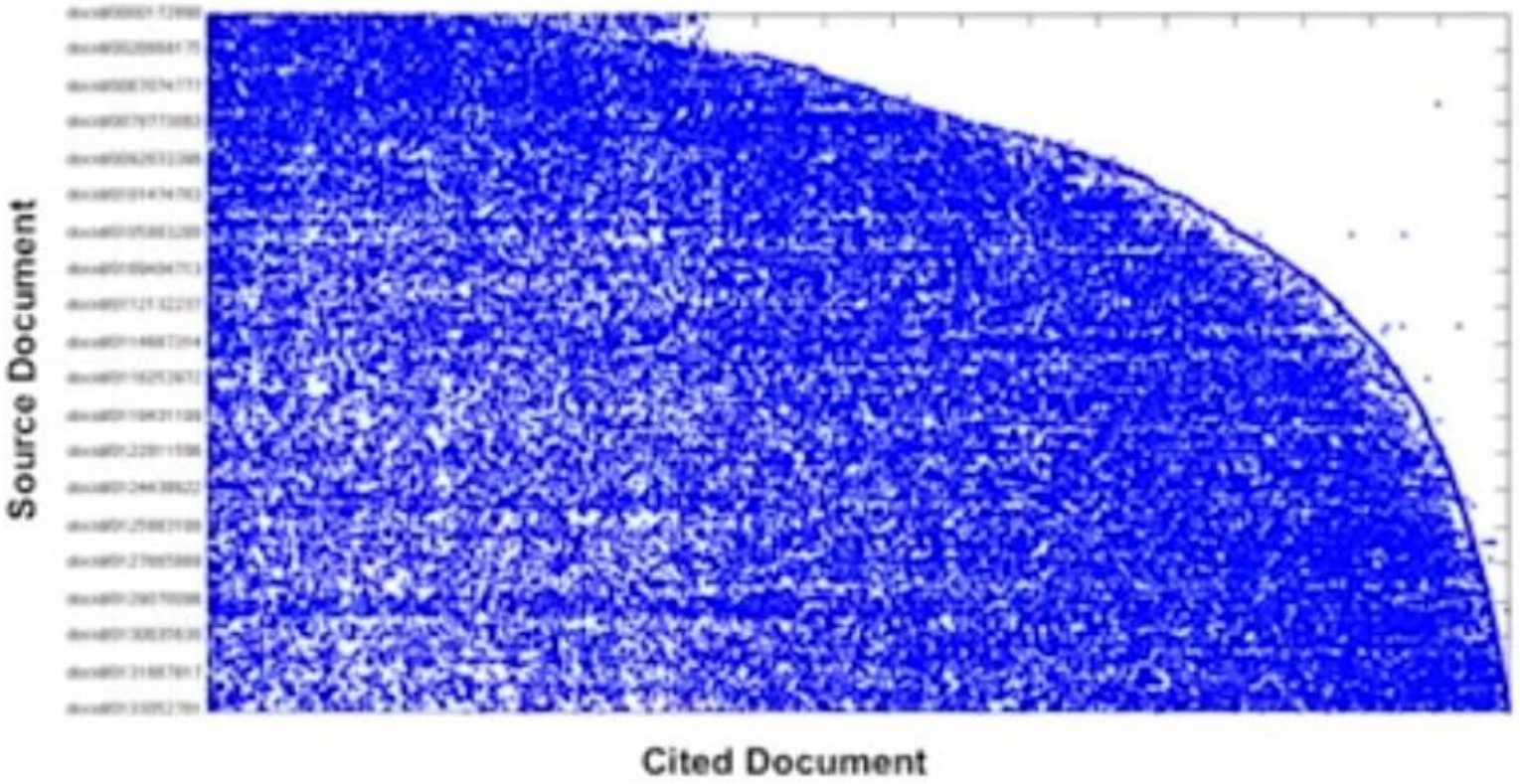
1. Uncompress XML file [~1 hour]
2. Read XML into binary structure and parse into triples [~2 hours]
3. Construct D4M associative arrays from triples to check data [~1 day]
4. Insert triples into Accumulo [key ~2 days, txt ~1 day, ngram ~10 days]

- Single node sustained insert rate of 10,000 – 100,000 entries/sec.
- Performance is sufficient that entire data set can be hosted on a single node

Outline

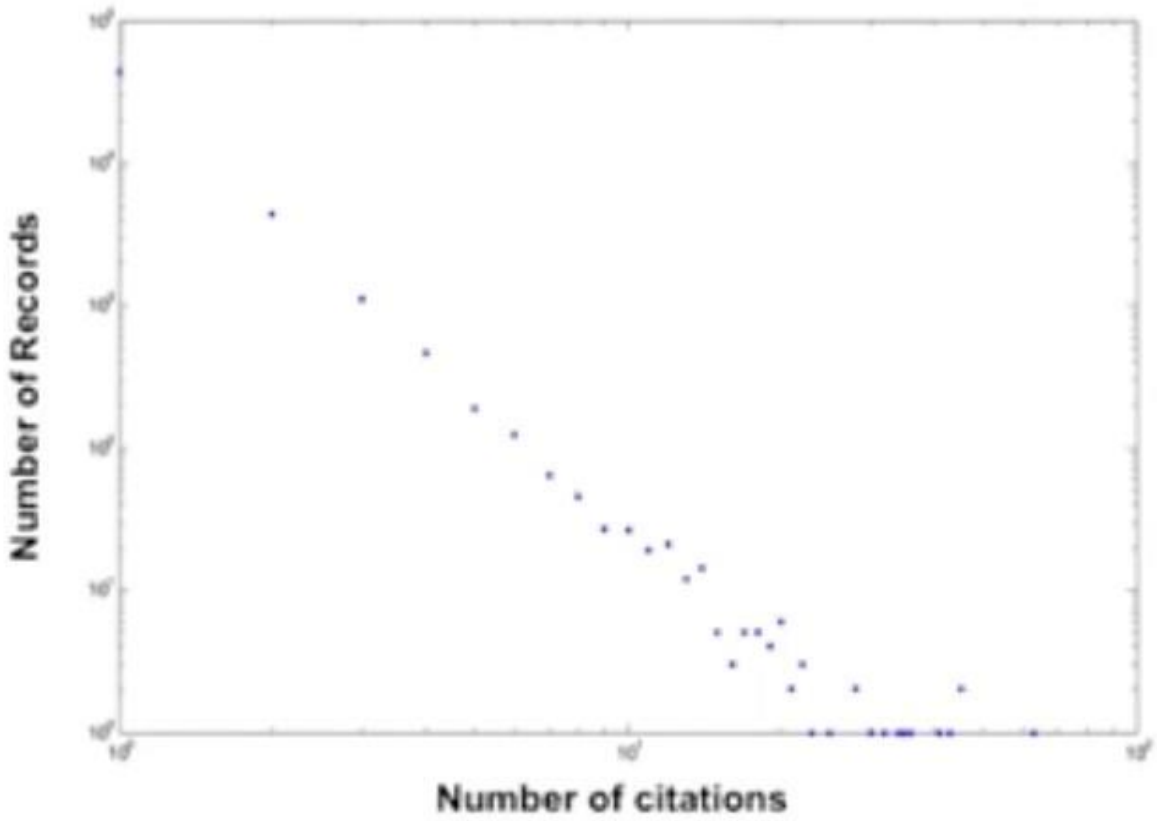
- Citation Data
- Graph Construction
  - Citation
  - Author
  - Institution
  - Keyword
  - Uncertainty
  - Pedigree
- Multi-Hyper Graphs
- Summary

Adjacency Matrix



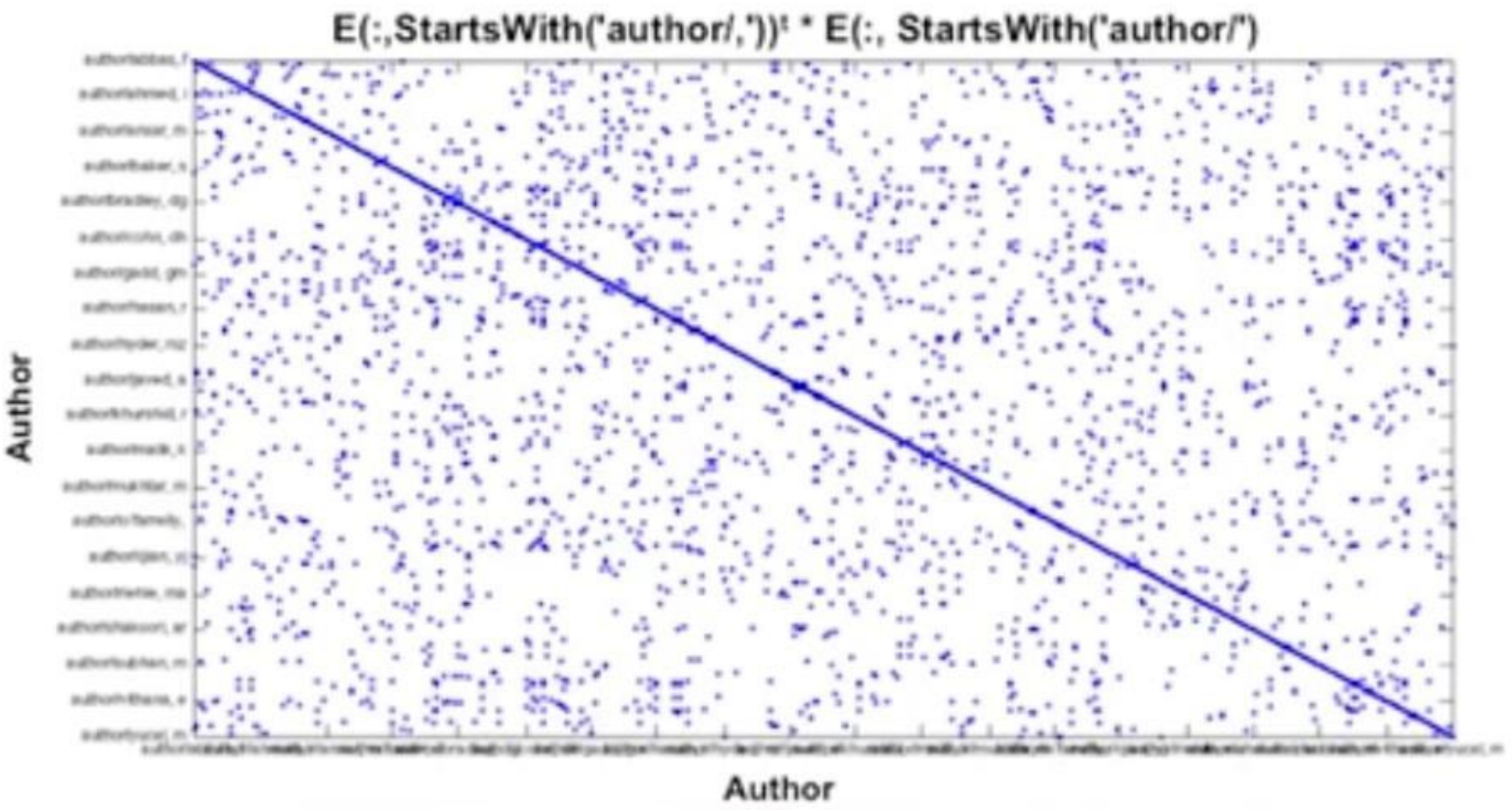
- Document ID increases with time (as expected)

Degree Distribution



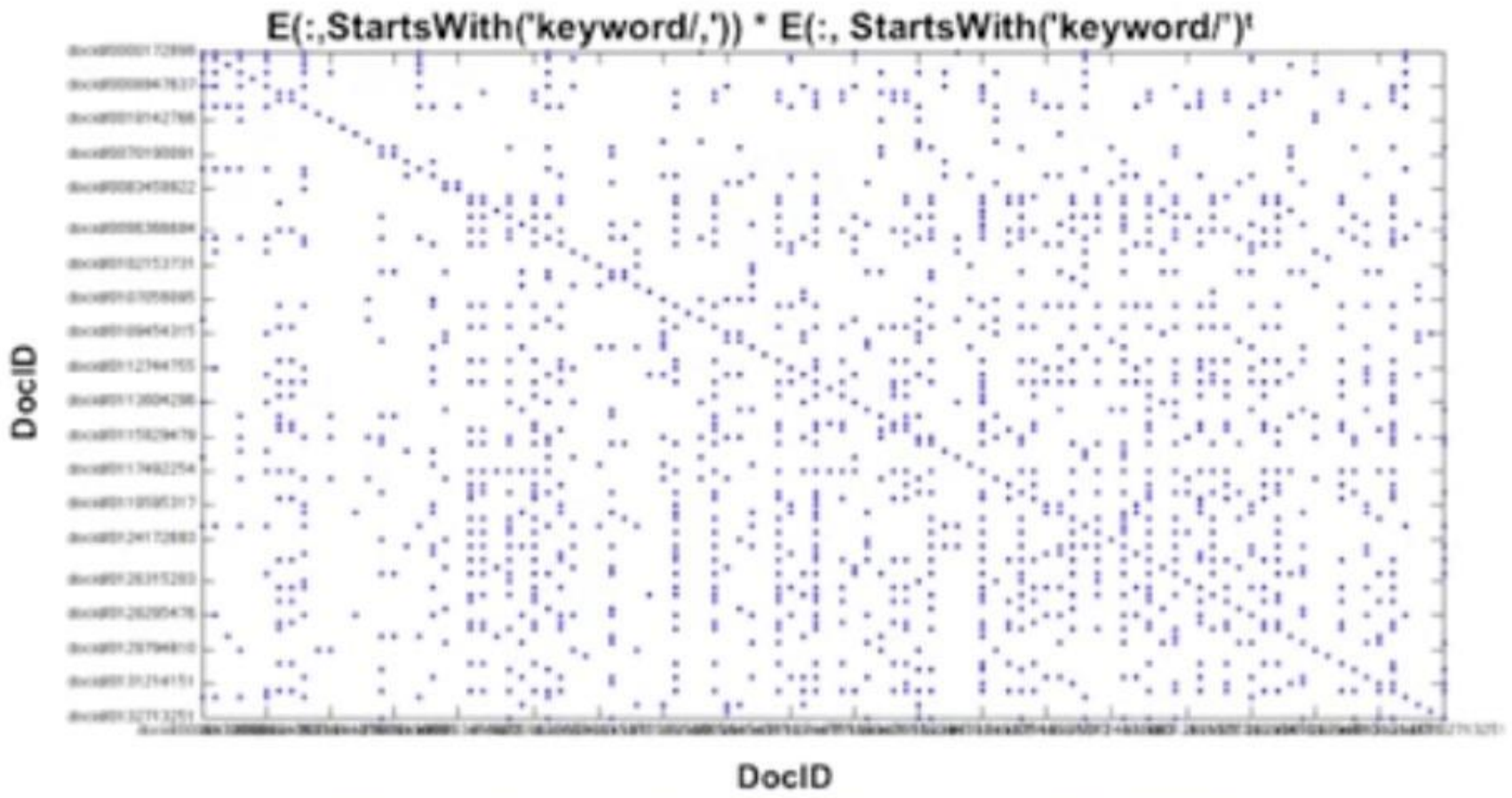
- Power law (as expected)

Author Graph



- Counts how many times a pair of Authors are in the same DocID

Keyword DocID Graph

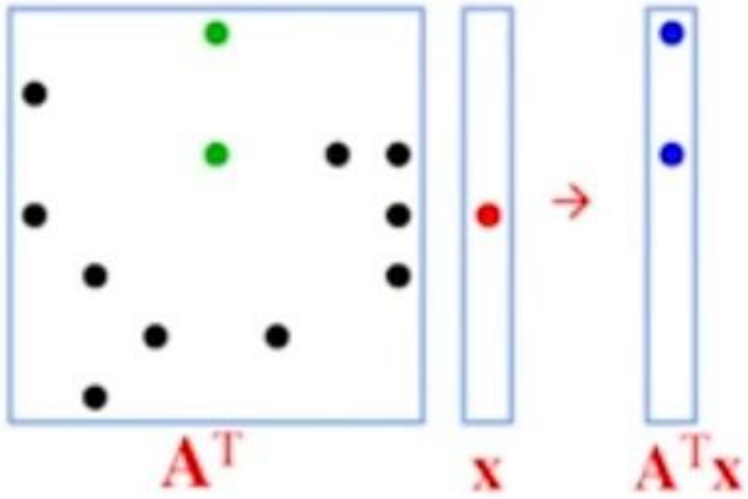


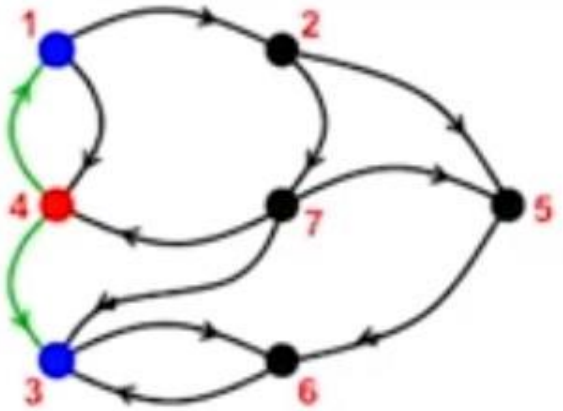
- Counts how many times a pair of DocIDs share a Keyword

Outline

- Citation Data
- Graph Construction
- Multi-Hyper Graphs
  - Undirected
  - Directed
  - Multi
  - Hyper
- Summary

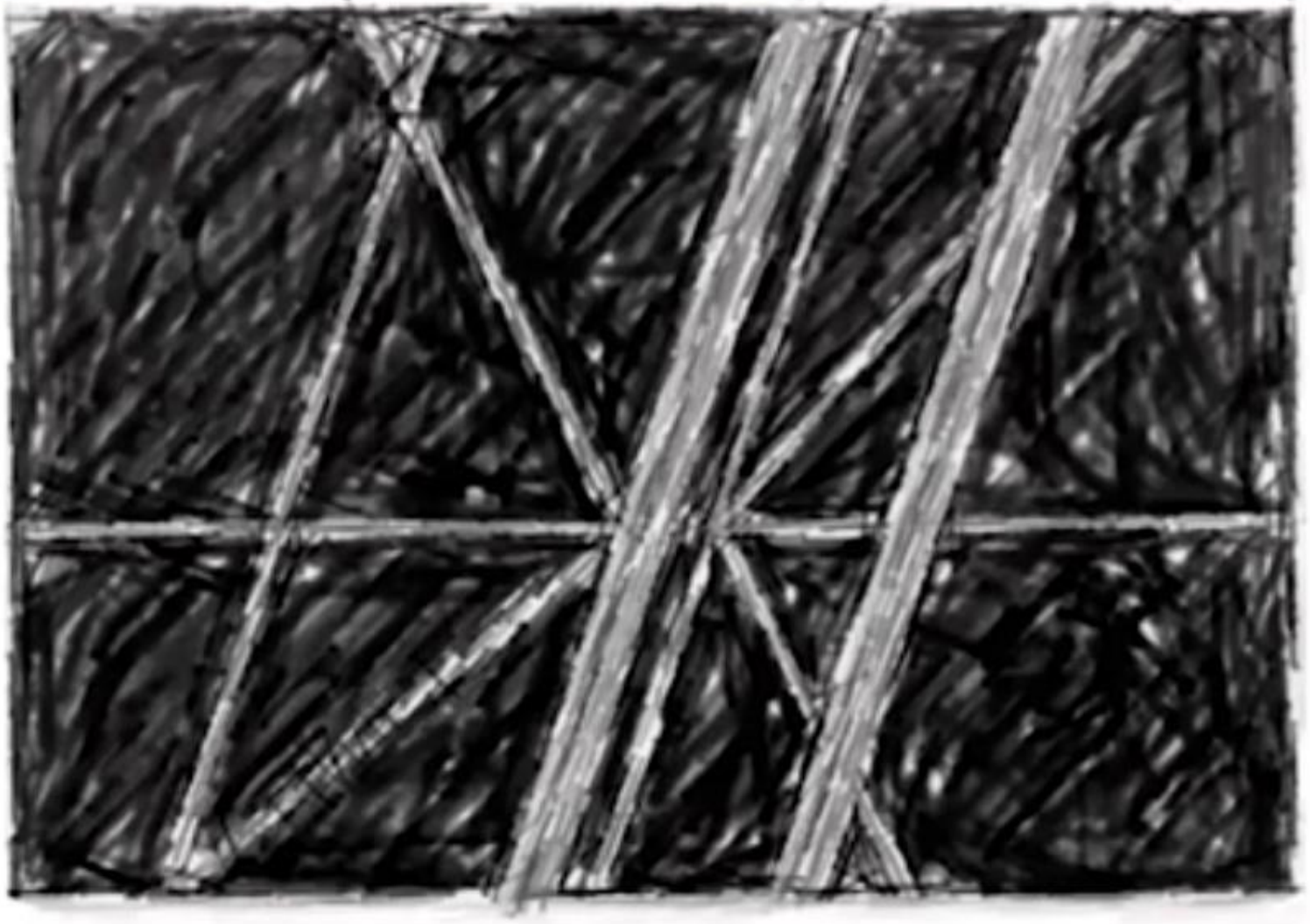
Directed Graph





- Directed graphs can be represented as a sparse matrices
  - Multiply by adjacency matrix – step to neighbor vertices
  - Work-efficient implementation from sparse data structures
- The real world is far more complex than directed graphs
  - Directed, multi, hypergraphs

Digraphs are Black & White



The World is Color

Artist: Ann Pibal; Painting: "XCRS"

Summary Observations

- Standard edge representation fragments hyper edges
  - Information is lost
- Digraph representation compresses multi-edges
  - Information is lost
- Matrix representation drops edge labels
  - Information is lost
- Standard graph representation drops edge order
  - Information is lost
- Need edge representation that preserves information

Artist: Ann Pibal; Painting: "XCRS"

Solution: Incidence Matrix

Edge	Color	Order	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
B1	Blue	2	1	1	1																	
S1	Silver	2	1	1	1																	
G1	Green	2	1	1	1																	
O1	Orange	2	1	1	1																	
O2	Orange	2	1	1	1																	
P1	Pink	2	1	1	1																	
B2	Blue	2				1	1	1	1	1												
S2	Silver	2				1	1	1	1	1												
G2	Green	2				1	1	1	1	1												
O3	Orange	2				1	1	1	1	1												
O4	Orange	2				1	1	1	1	1												
P2	Pink	2				1	1	1	1	1												
O5	Orange	1		1				1			1		1					1				1
P3	Pink	2										1	1		1	1						
P4	Pink	2		1								1										
P5	Pink	2		1										1	1							
P6	Pink	2															1	1	1			
P7	Pink	3					1										1			1		
P8	Pink	3							1									1			1	

Artist: Ann Pibal; Painting: "XCRS"