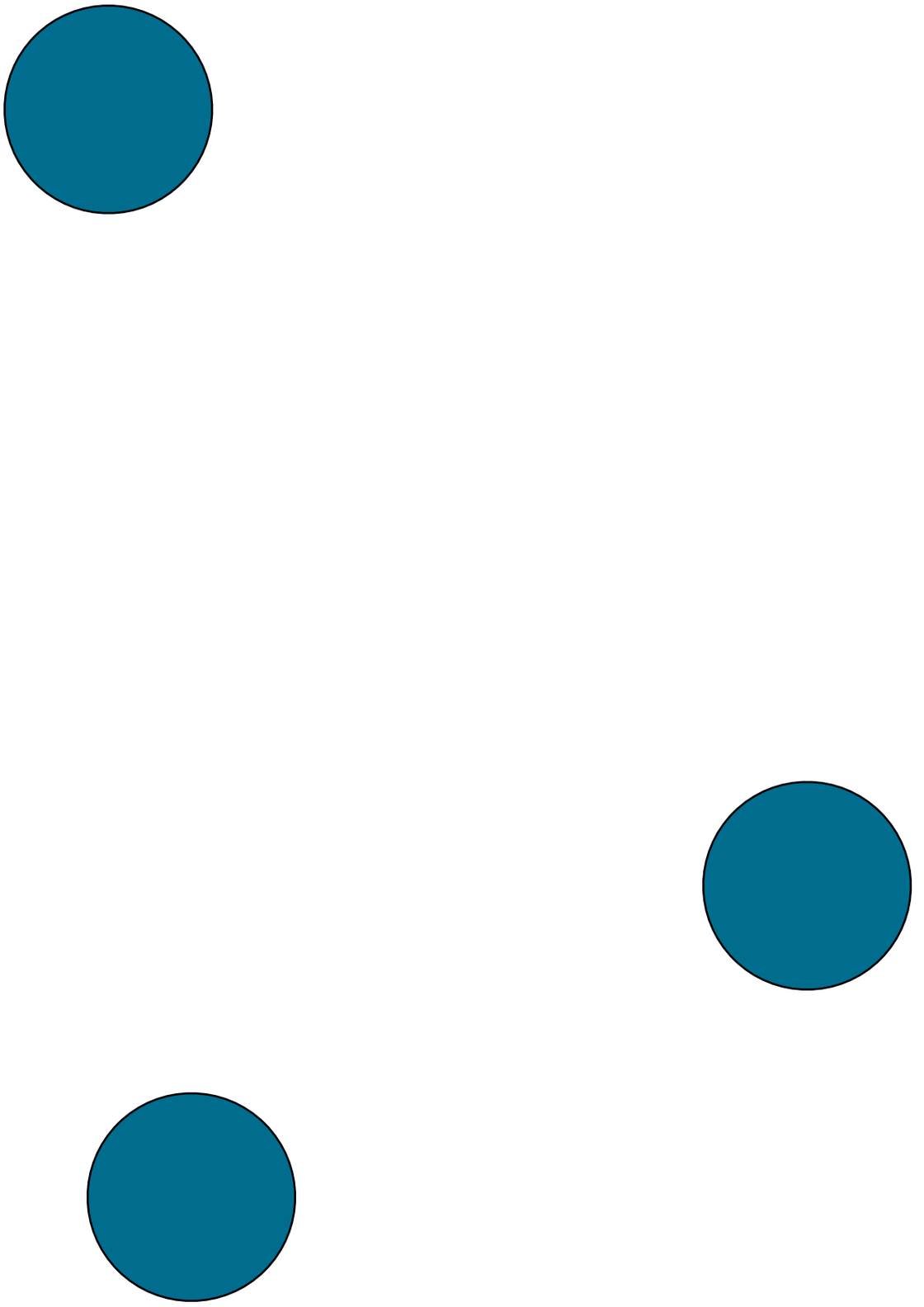
Structure between data sources

David Bamman
Info 202: Information Organization and Retrieval

October 17, 2016

Networks

Network	Nodes	Edges	Information
	People		
	Servers		
	Articles		
	Web pages		



Summary: centrality

What's important?	Measure
Number of friends	Degree centrality
Number or importance of friends	Eigenvector, Katz centrality; PageRank
Distance from others	Closeness centrality
Middleman	Betweenness centrality

$$\text{Degree(i)} = \sum_j A_{i,j}$$

	1	2	3	4	5
1			1		
2			1		1
3	1	1		1	1
4			1		
5		1	1		

Degree
1
2
4
1
2

Centrality

- Eigenvector centrality

$$\textit{centrality}(i) = \sum_j \left[A_{i,j} \times \textit{centrality}(j)\right]$$

- Katz centrality

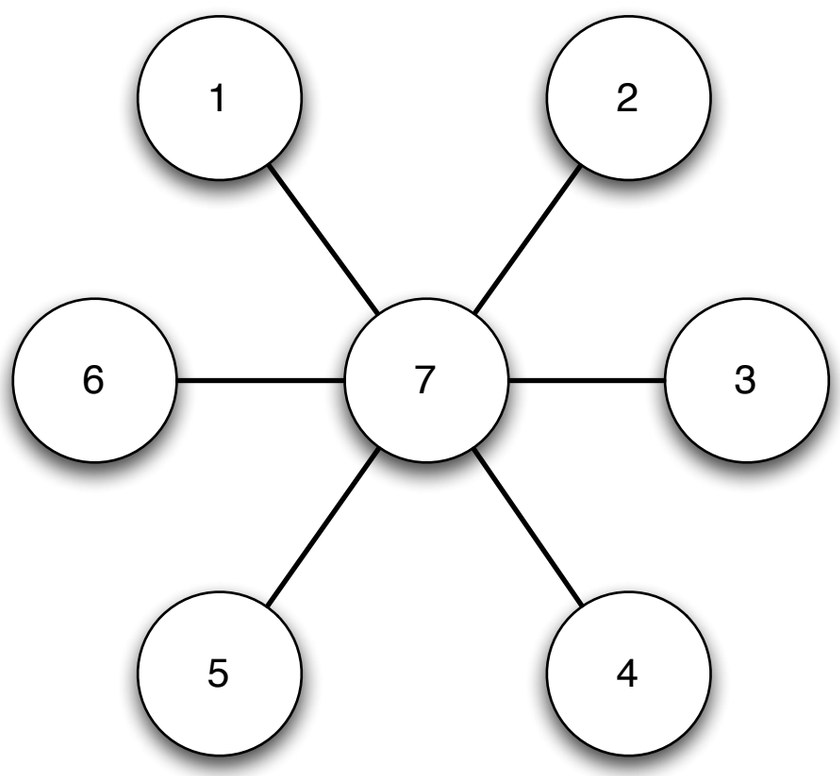
$$\textit{centrality}(i) = \alpha \times \sum_j \left[A_{i,j} \times \textit{centrality}(j)\right] + \beta$$

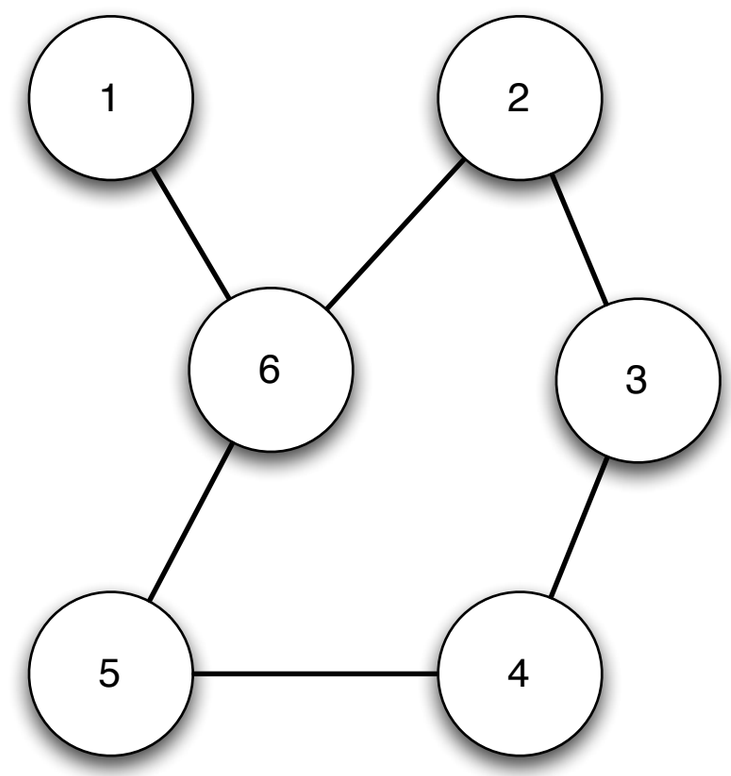
- PageRank

$$\textit{centrality}(i) = \alpha \times \sum_j \left[A_{i,j} \times \frac{\textit{centrality}(j)}{\textit{outdegree}(j)}\right] + \beta$$

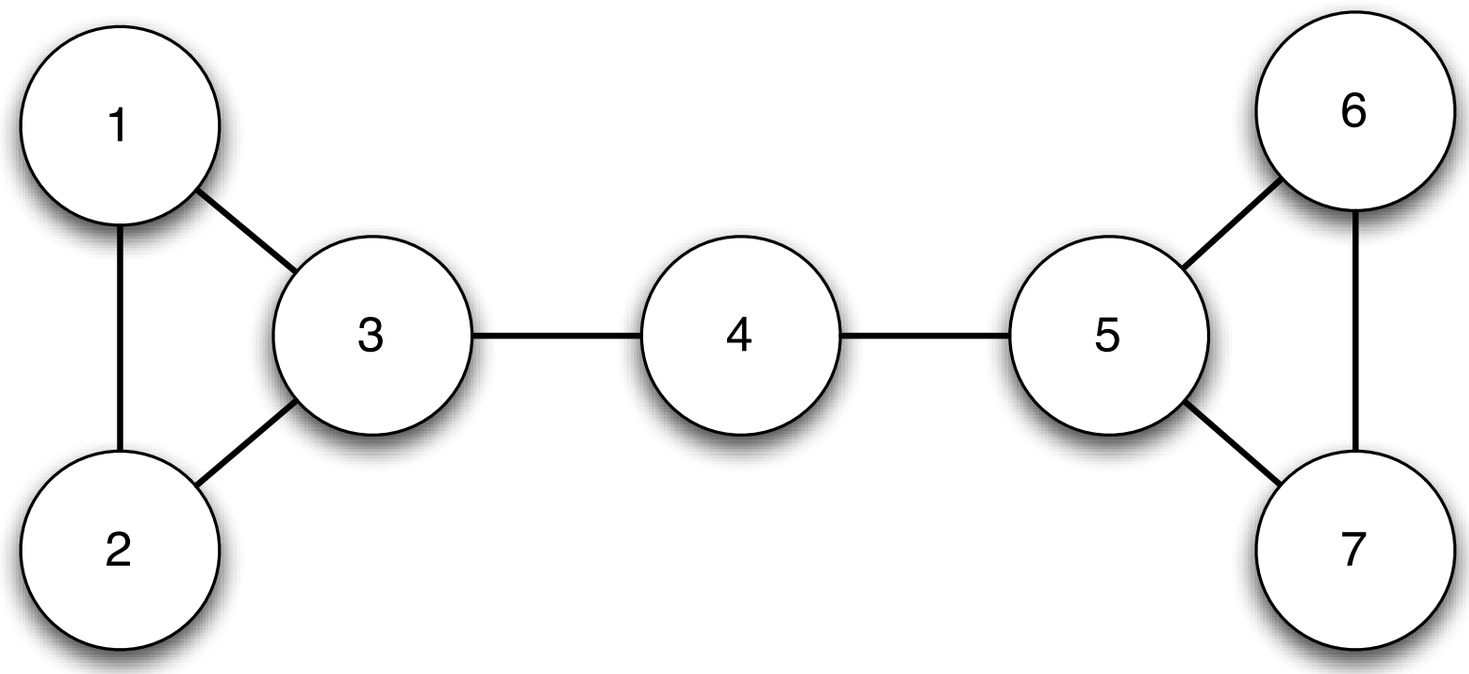
Closeness centrality

$$centrality(i) = \frac{\sum_j shortest\_path(i, j)}{n}$$



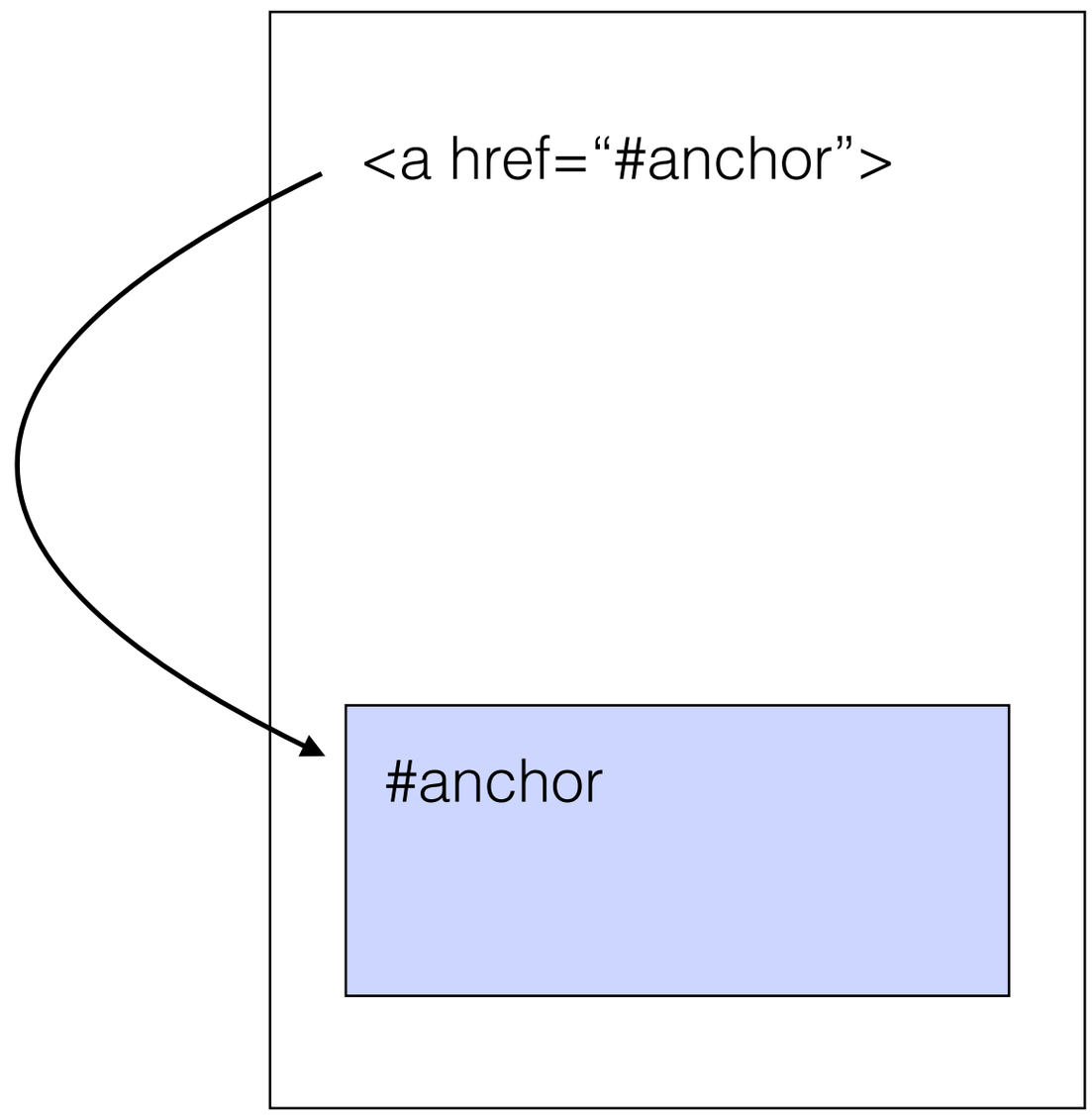


Betweenness centrality



$$betweenness(i) = \sum_{s,t} \mathrm{I}\{i \text{ is on the path from s to t}\}$$

Structure within a resource



#anchor

Table of Contents

Table of Contents

Chapter 1
Foundations for Organizing Systems

Chapter 2
Design Decisions in Organizing Systems

Chapter 3
Activities in Organizing Systems

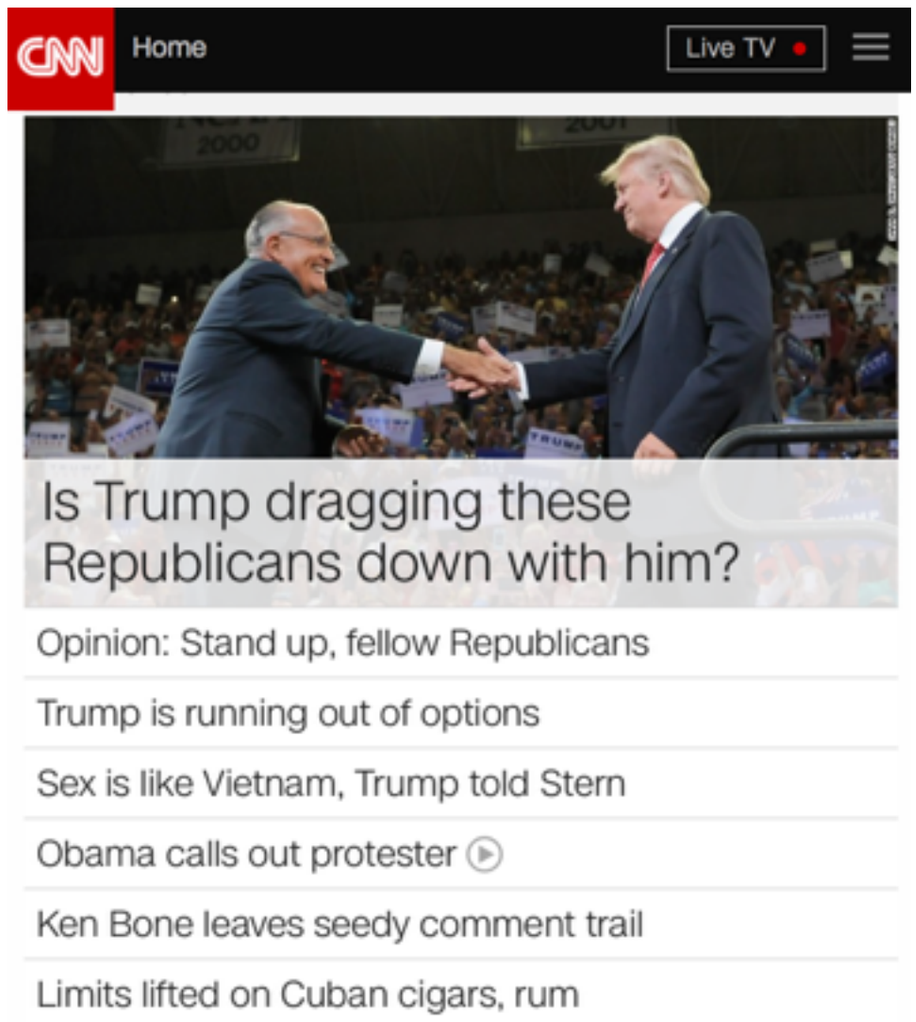
Index

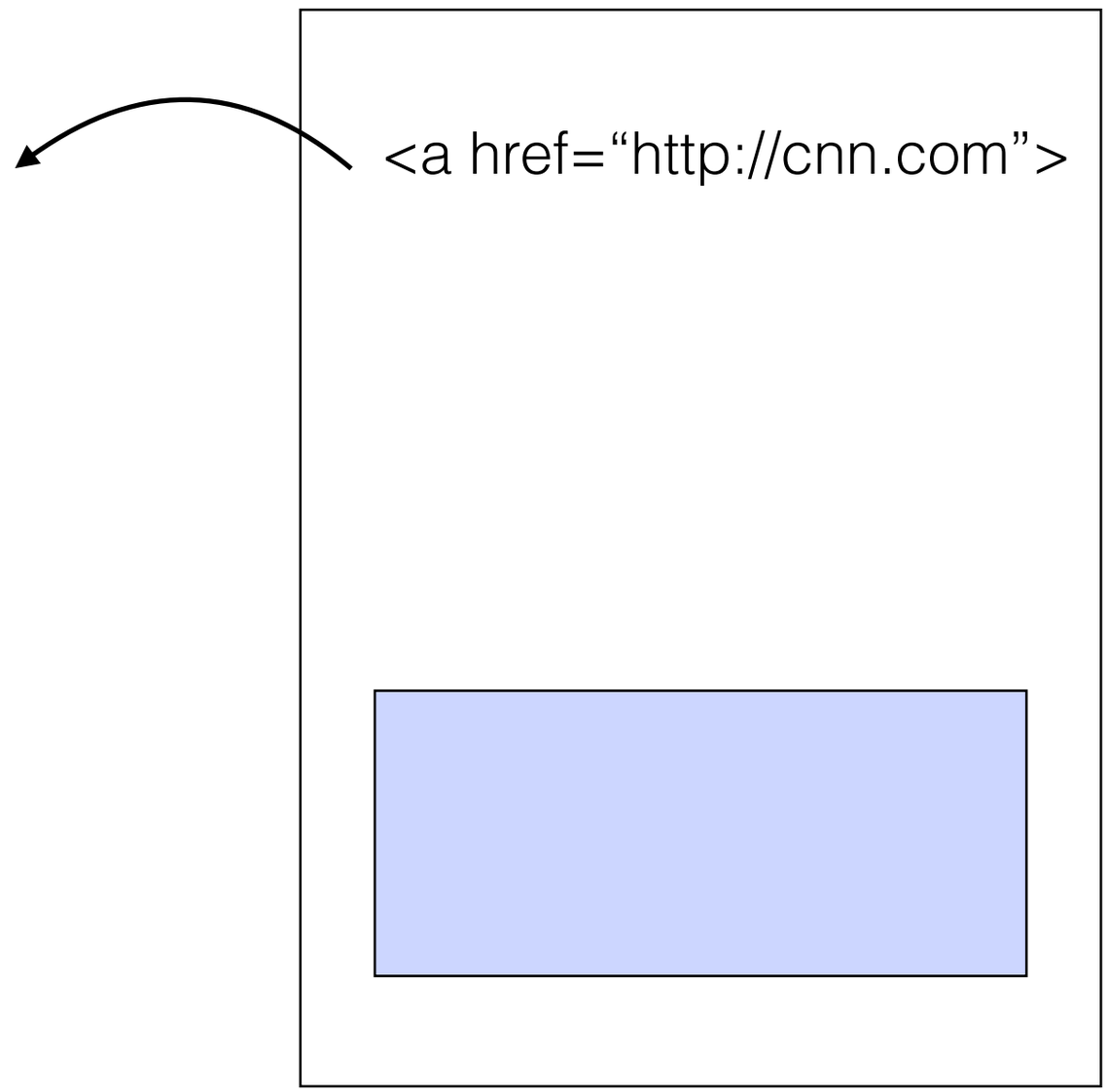
Chapter 1
Foundations for Organizing Systems

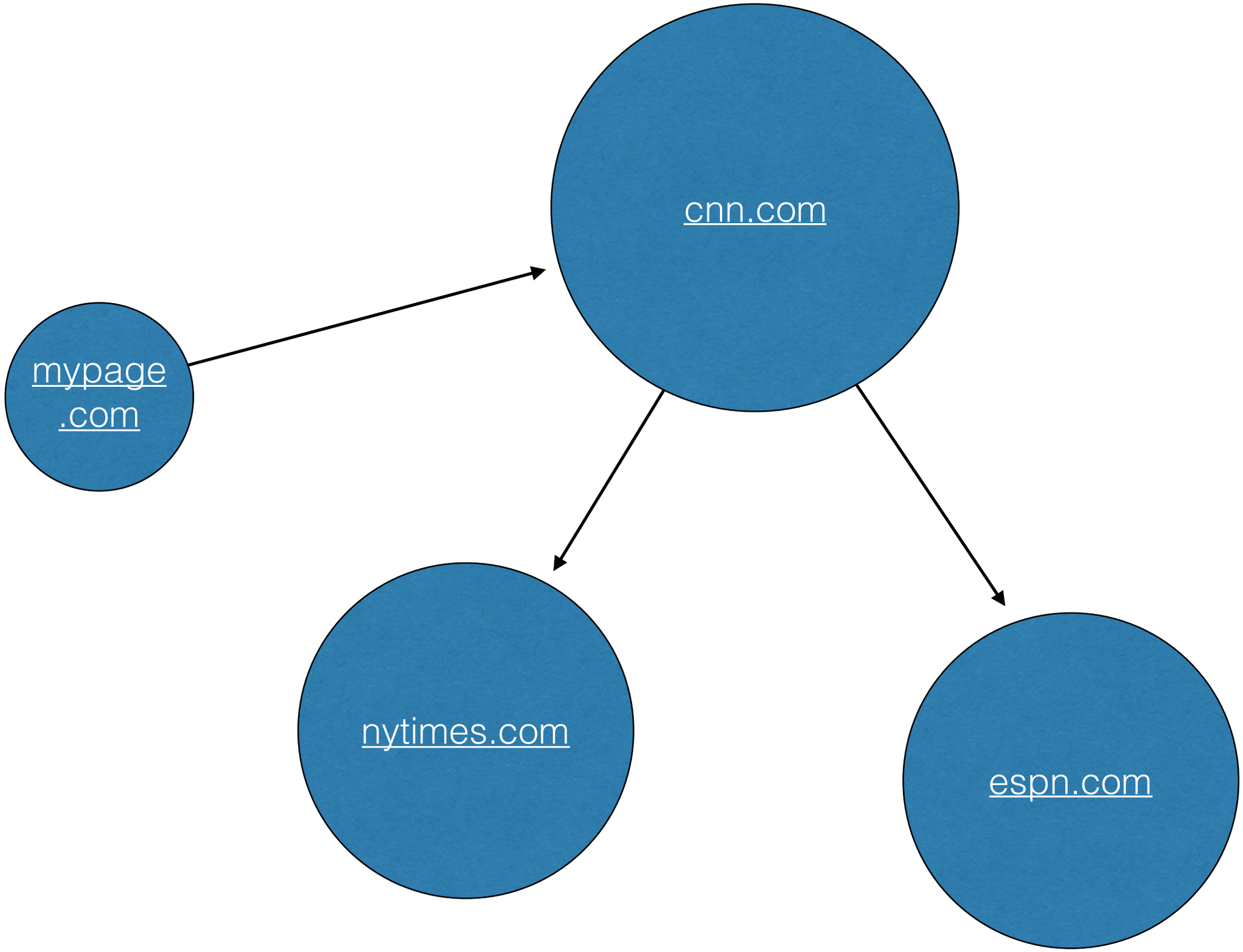
Chapter 2
Design Decisions in Organizing Systems

Chapter 3
Activities in Organizing Systems

Structure between resources






cnn.com
mypage
.com
nytimes.com
espn.com

Annotations

Search lyrics & more

GENIUS

SIGN UP SIGN IN

HOME LYRICS TEXTS FORUMS ADD SONG NEWS GENIUS ANNOTATE THE WEB SHOP

61 **Triumph**

Wu-Tang Clan

Featuring Cappadonna
Produced By RZA
Album Wu-Tang Forever

[Verse 1: Inspectah Deck]
I bomb atomically, Socrates' philosophies and hypotheses
Can't define how I be dropping these mockeries
Lyrically perform armed robbery
Flee with the lottery, possibly they spotted me
Battle-scarred Shogun, explosion when my pen hits
tremendous

Genius Annotation 6 contributors

“ Socrates taught that when you're searching for knowledge, you being by defining all known terms. Deck's saying his disses defy the comprehension of even the most complete set of philosophical knowledge. Source: The Wu-Tang Manual (2005)

Upvote +86 Share

Annotations

Bibliometrics

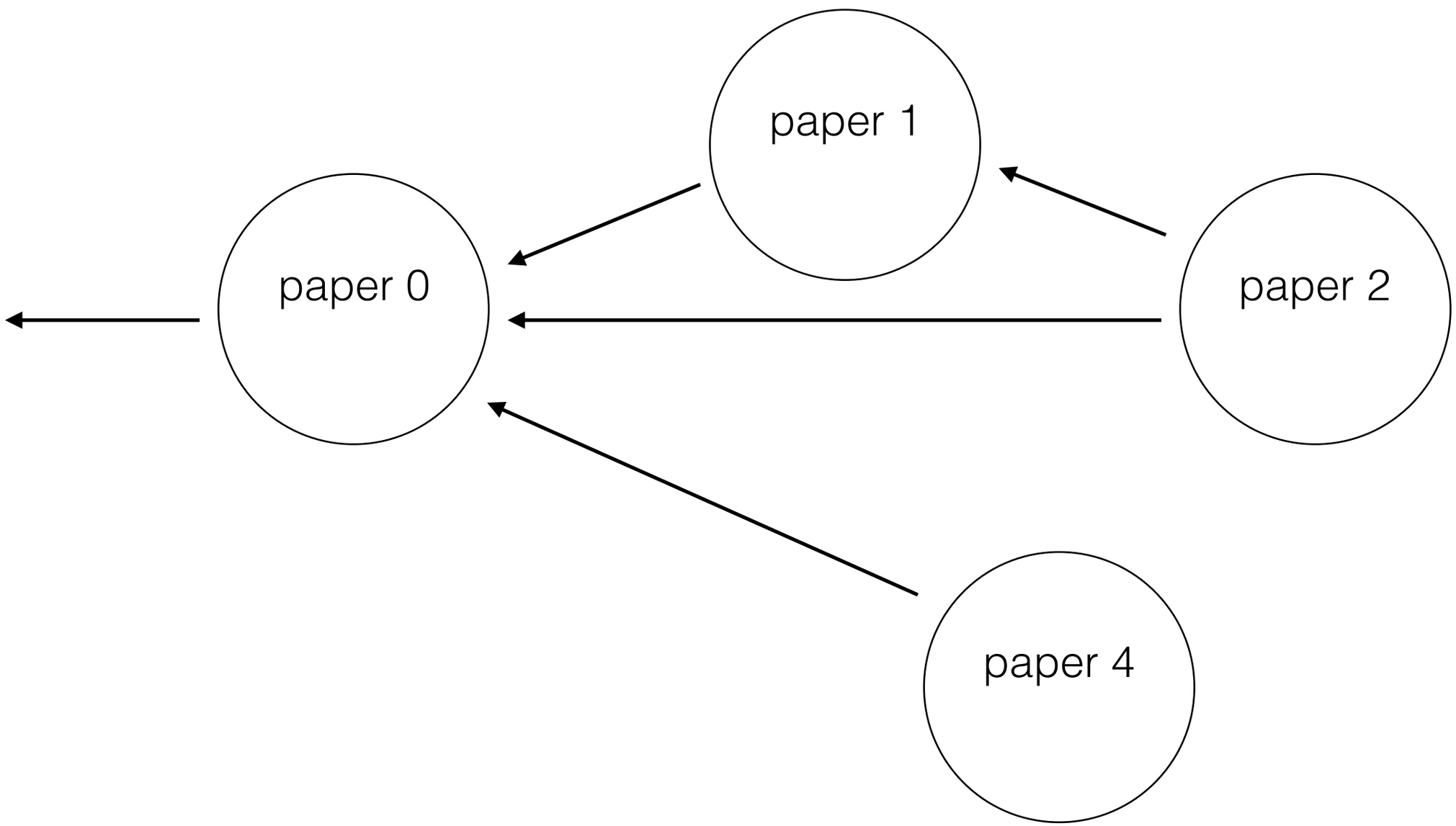
- Analysis of scientific citation began in the 1920s as a way to quantify the influence of specific documents or authors in terms of their "impact factor"

- It can also identify "invisible colleges" of scientists whose citations are largely self-referential

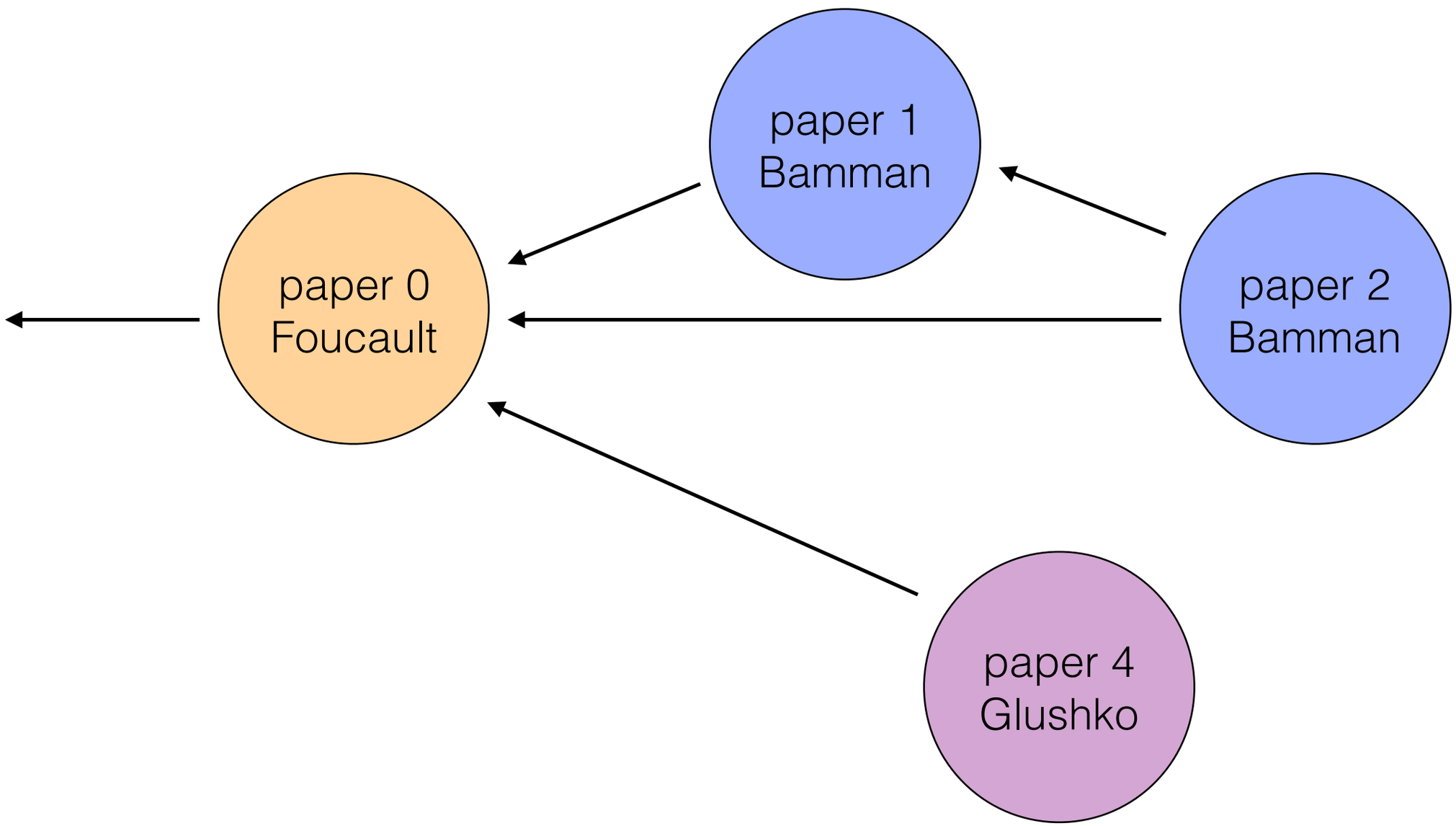
- It can recognize the emergence of new scientific disciplines

- (Eugene Garfield and Derek J. de Solla Price are two of the "founding fathers" of bibliometrics)

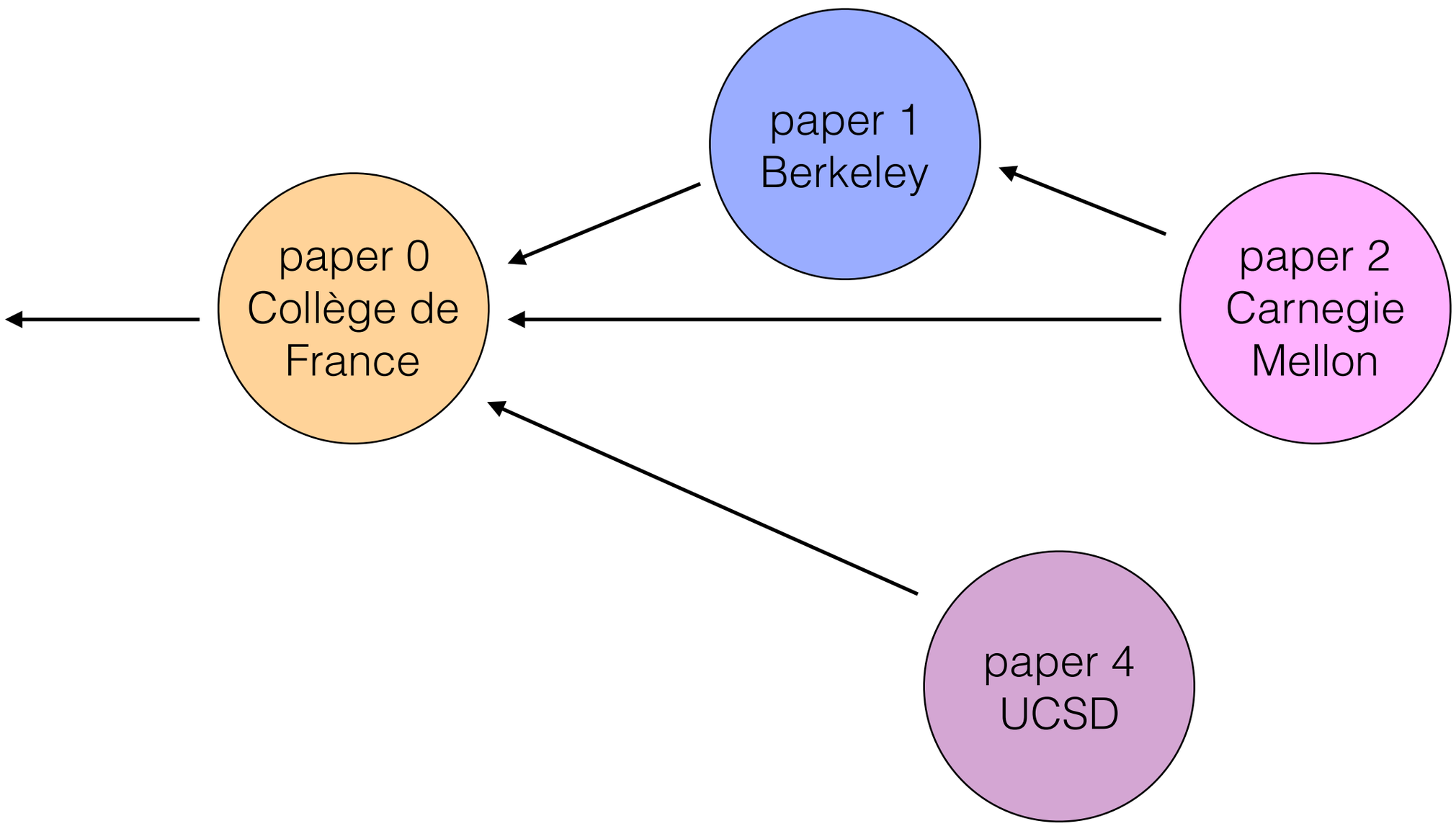
Metrics



Metrics


paper 1
Bamman
paper 0
Foucault
paper 2
Bamman
paper 4
Glushko

Metrics


paper 1
Berkeley
paper 0
Collège de
France
paper 2
Carnegie
Mellon
paper 4
UCSD

Citation polarity

- When one resource cites another there is often a lexical signal that indicates how a writer views the relationship of a citation to the text from which the citation is made

- A citation or link without a signal suggests by default that the citation supports the current text

- Explicit signals that indicate positive polarity include "See," "See also," "See generally," and "Cf."

- Signals that indicate negative polarity include "But see" and "Contra"

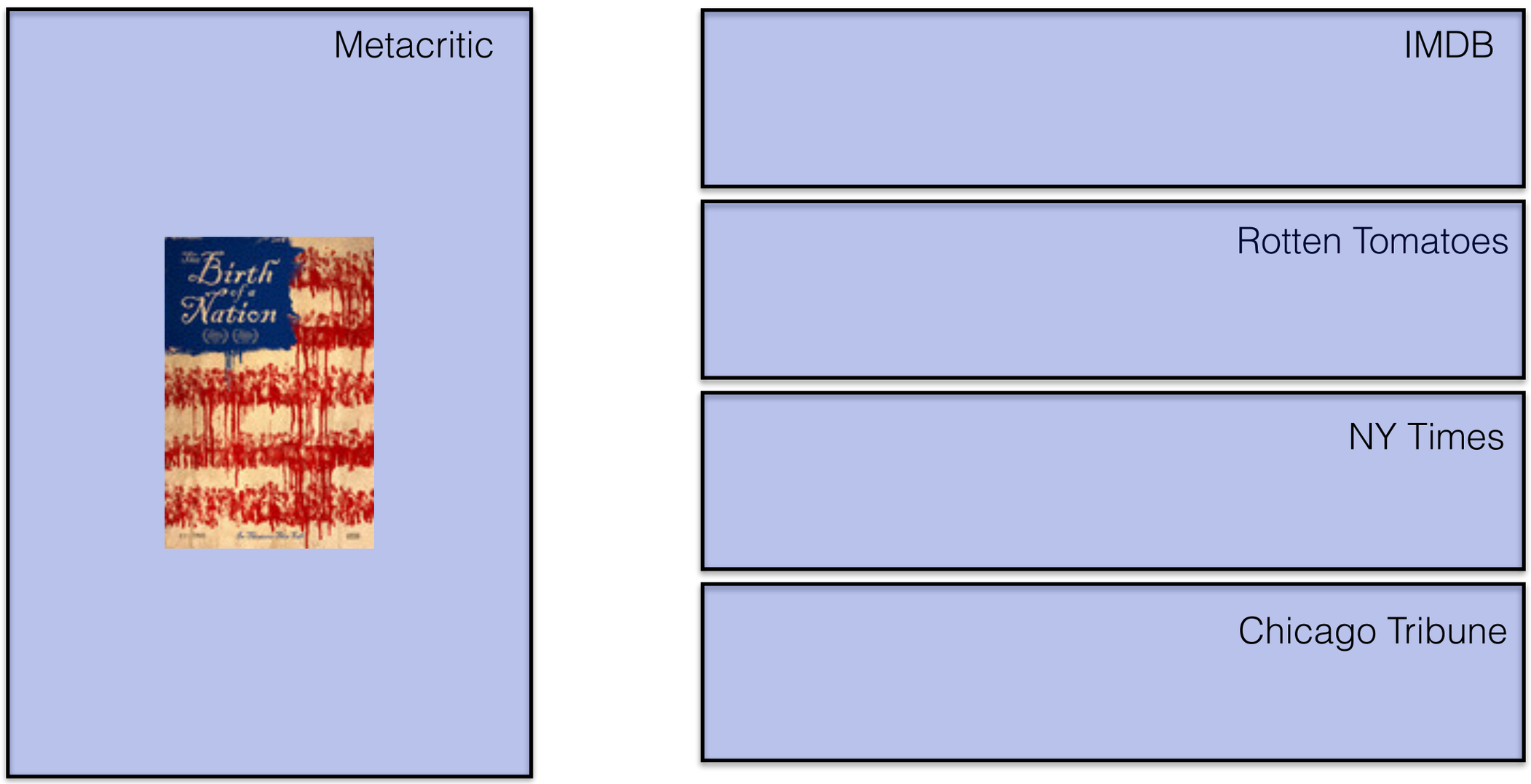
Altmetrics

- People are lazy
- Most papers cite only a small proportion of the sources that influenced them
- Secondary sources are cited more than primary ones, because people don't know the literature, and informal ones aren't cited
- People cite their friends and themselves more than is justified

Altmetrics

- The “Altmetrics” movement is trying to make non-traditional contributions count for academic evaluations
- Publishing in “open publications” – measuring downloading
- Sharing “raw science” like datasets, code, and experimental designs
- Blogging, microblogging, and comments or annotations on existing work

Structure between resources



One solution: the semantic web

The Shining

Resource URI: http://data.linkedmdb.org/resource/film/2014

Home | Example film

Property	Value
movie:actor	<http://data.linkedmdb.org/resource/actor/29704>
movie:actor	<http://data.linkedmdb.org/resource/actor/30013>
movie:actor	<http://data.linkedmdb.org/resource/actor/33144>
movie:actor	<http://data.linkedmdb.org/resource/actor/35070>
movie:actor	<http://data.linkedmdb.org/resource/actor/39390>
movie:actor	<http://data.linkedmdb.org/resource/actor/44448>
movie:actor	<http://data.linkedmdb.org/resource/actor/45066>
movie:actor	<http://data.linkedmdb.org/resource/actor/45772>
movie:actor	<http://data.linkedmdb.org/resource/actor/47299>
movie:actor	<http://data.linkedmdb.org/resource/actor/60994>
movie:actor	<http://data.linkedmdb.org/resource/actor/60995>
movie:actor	<http://data.linkedmdb.org/resource/actor/8971>
movie:actor	<http://data.linkedmdb.org/resource/actor/8987>
foaf:based_near	<http://sws.geonames.org/2635167/>
movie:country	<http://data.linkedmdb.org/resource/country/GB>
dc:date	1980-05-23
movie:director	<http://data.linkedmdb.org/resource/director/8476>
movie:editor	<http://data.linkedmdb.org/resource/editor/2881>
movie:editor	<http://data.linkedmdb.org/resource/editor/88>
movie:featured_film_location	<http://data.linkedmdb.org/resource/film_location/318>
movie:featured_film_location	<http://data.linkedmdb.org/resource/film_location/422>
movie:featured_film_location	<http://data.linkedmdb.org/resource/film_location/772>
movie:featured_film_location	<http://data.linkedmdb.org/resource/film_location/803>
movie:featured_film_location	<http://data.linkedmdb.org/resource/film_location/990>

Linking databases

- Many of the same resources are described in different datasets in different ways
    - different levels of granularity
    - different properties
    - different perspectives on the values of those properties
- Many opportunities for inference if you can link the same entities across datasets

Database linking

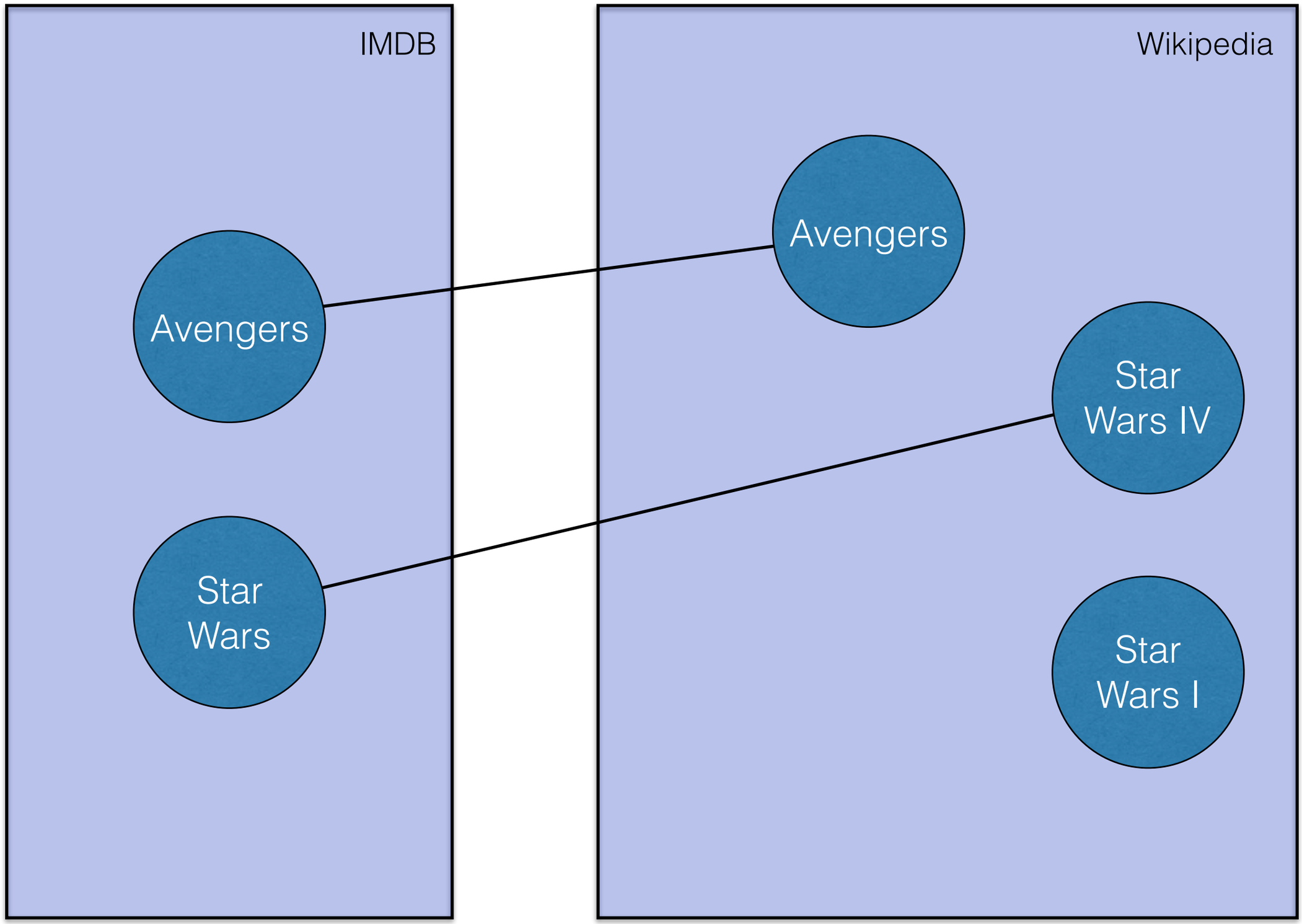
- Voting records to the deceased
- Documents from SEC, Pentagon, defense contractors to note movement to industry (Cohen 2012)
- Press releases from different members of congress
- Indictments/settlements from U.S. attorneys
- DSA database of safety status of CA public schools + US seismic zones + school list from CA Dept of Education (Parasie 2015)

IMDB

Movie	Rating
The Avengers	4
Star Wars	5

Wikipedia

Movie	Box office
The Avengers	1.52B
Star Wars IV: A New Hope	775M
Star Wars I: The Phantom Menace	1.03B


IMDB
Avengers
Star
Wars
Wikipedia
Avengers
Star
Wars IV
Star
Wars I

Entity linking

- Email contacts database
- Academic publications
- Book catalogues

Data science lifecycle

Cross Industry Standard Process for Data Mining (CRISP-DM)

Similarity

- We'll talk on Wednesday about the conceptual foundations of similarity and how it's used for categorization → classification
- We can exploit similarity here as well for the related task of ranking entities

Feature description

IMDB

Movie	Rating
The Avengers	4
Star Wars	5
Rocky 4	3
Rocky	5

Wikipedia

Movie	Box office
Marvel's Avengers	$1.52B
Star Wars IV: A New Hope	$775M
Star Wars I: The Phantom Menace	$1.03B
Rocky	$225M
Rocky IV	$300M

Bag of unigrams

- Represent a string as the count of tokens within it

	the avengers	marvel's avengers	rocky
the	1		
avengers	1	1	
star			
wars			
rocky			1
4			
marvel's		1	
IV			
a			
new			
hope			
phantom			
menace			

Jaccard Similarity

	the avengers	marvel's avengers
the	1	
avengers	1	1
star		
wars		
rocky		
4		
marvel's		1

number of features in **both** X and Y

$$\frac{|X \cap Y|}{|X \cup Y|}$$

number of features in **either** X and Y

Bag of character ngrams

- Represent a string as the count of sequences of characters of length *n*.

	the avengers	marvel's avengers	rocky
the	1		
aven	1	1	
veng	1	1	
enge	1	1	
nger	1	1	1
gers	1	1	
marv		1	
arve		1	
rvel		1	
vel'		1	
el's		1	
ls'		1	
s' a		1	
…			

Weighting

Some tokens/ngrams show up much more frequently in titles (and are hence less informative for similarity

Movie	Rating
The Avengers	4
Star Wars	5
Rocky 4	3
Rocky	5

Movie	Box office
Marvel's Avengers	$1.52B
Star Wars IV: A New Hope	$775M
Star Wars I: The Phantom Menace	$1.03B
Rocky	$225M
Rocky IV	$300M

TF-IDF

- Term frequency-inverse document frequency
- A scaling to represent a feature as function of how frequently it appears in a data point but accounting for its frequency in the overall collection
- IDF for a given term = the number of documents in collection / number of documents that contain term

Cosine Similarity

$$cos(x, y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2}\sqrt{\sum_{i=1}^{F} y_i^2}}$$

- Jaccard similiarty is measure of set overlap.
- Cosine similarity reasons over the value of features
- Often weighted by TF-IDF to discount the impact of frequent features.

String tranformations

- Misspellings (and other subtle transformations) are costly for measures that reason about the identify of feature

Movie	Rating
The Avengers	4

Movie	Box office
Marvel's Avengrs	$1.52B

Token jaccard similarity = 0

Edit distance

- Edit distance = the similarity between two strings based on the minimal number of additions, deletions and substitutions it takes to get from one to the other

word 1	word 2	edit distance	
avengers	avegrs	1	deletion
avengers	avengeers	1	addition
avengers	evengers	2	substitution
avengers	car	11	additions + substitions

edit distance based for costs: addition = 1, deletion = 1, substitution = 2

Complex representations

IMDB

Movie		Rating
The Avengers	2012	4
Star Wars	1977	5
Rocky 4	1985	3
Rocky	1976	5

Wikipedia

Movie		Box office
Marvel's Avengers	2012	$1.52B
Star Wars IV: A New Hope	1978	$775M
Star Wars I: The Phantom Menace	1999	$1.03B
Rocky	1976	$225M
Rocky IV	1985	$300M

Complex representations

First name	M.I	Last name
Jon		Snow
Jonathan	I.	Snow
John	I.	Snow
Robert	I.	Snow
Jon	F.	Snow

	jon snow	jonathan i. snow	john i. snow
FN:jon	1	1	
FN:ona		1	
FN:nat		1	
FN:ath		1	
MI:∅	1		1
MI:I.		1	1
MI:F.			
LN:snow	1	1	1

Structure

- How does this relate to network analysis?

Moving to IR

- Identifying duplicate entities in a database (and matching entities between databases) starts moving us into information retrieval

Duplicate detection

PRESIDENT OBAMA MAKES HIS FINAL 4 PICKS; KANSAS AS CHAMPS

WASHINGTON (AP) -- President Barack Obama has made his final NCAA Tournament call in office: Rock Chalk, champions.

Obama picked Kansas, Texas A&M, North Carolina and Michigan State to all reach the Final Four in a bracket he filled out for ESPN.

AP Photo/Pablo Martinez Monsivais

President Obama Makes His Final 4 Picks
abcnews.go.com/.../president-obama-makes-final-picks
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama picks KU basketball as c
m.kusports.com/.../president-obama-picks-ku-basketba
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

WKTV.com | President Obama makes his
www.wktv.com/.../President_Obama_makes_his_Final
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama makes his Final 4 picks;
www.kswo.com/.../president-obama-makes-his-final-4-
His choice might be an unpopular one around Kansas,
correctly predicted the national champion since he picked

President Obama calls for Rock Chalk Cha
www.wibw.com/.../President-Obama-calls-for-Rock-Cha
2 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

President Obama makes his Final 4 picks;
https://www.artesianews.com/.../president-obama-mak
5 days ago - **His choice might be an unpopular one arou**
hasn't correctly predicted the national champion since he p

Duplicate document detection

- What are the resources we're comparing?
- How do we describe each one?
- How do we measure "similarity"
- Evaluation?

Computational concerns

- Two sources of complexity:
- Dimensionality of the feature space (every document in represented by a vocabulary of 1M word) [minhashing]
- Number of documents in collection to compare (4.64 billion web pages) [locality sensitive hashing]

Text Reuse

We were many times weaker than his splendid, lacquered machine, so that I did not even attempt to outspeed him. O lente currite noctis equi! O softly run, nightmares!

Nabokov, *Lolita*

Text reuse detection

- What are the resources we're comparing?
- How do we describe each one?
- How do we measure "similarity"
- Evaluation?

Information retrieval

Google conrad heart of darkness

All Books Images Videos Shopping More

About 479,000 results (0.46 seconds)

Heart of Darkness - Wikipedia, the free encyclop
https://en.wikipedia.org/wiki/**Heart_of_Darkness** ▾ Wikipedi
Heart of Darkness (1899) is a novella by Polish-British novelist Jo
voyage up the Congo River into the Congo Free State, in the heart
Joseph Conrad - Kurtz - Disambiguation - Léon Rom

SparkNotes: Heart of Darkness
www.sparknotes.com/lit/**heart/** ▾ SparkNotes ▾
Heart of Darkness. Joseph **Conrad** ... Buy the print **Heart of Dar**
BN.com ... Order **Heart of Darkness** and Selected Short Fiction a
Part 1 - Part 2 - Part 3 - Context

Heart of Darkness, by Joseph Conrad - Project
www.gutenberg.org/files/219/219-h/219-h.htm ▾ Project Gute
The Project Gutenberg EBook of **Heart of Darkness**, by Joseph C
for the use of anyone anywhere at no cost and with almost no rest

Heart of Darkness - Shmoop
www.shmoop.com › Literature ▾
We really can't say it better than Joseph **Conrad** himself. **Heart of**
story of a journalist who becomes manager of a station in the (Afri

Heart of Darkness at a Glance - Cliffs Notes
www.cliffsnotes.com/.../**heart-of-darkness/heart-of-darkness**
Joseph **Conrad**'s **Heart of Darkness** retells the story of Marlow's

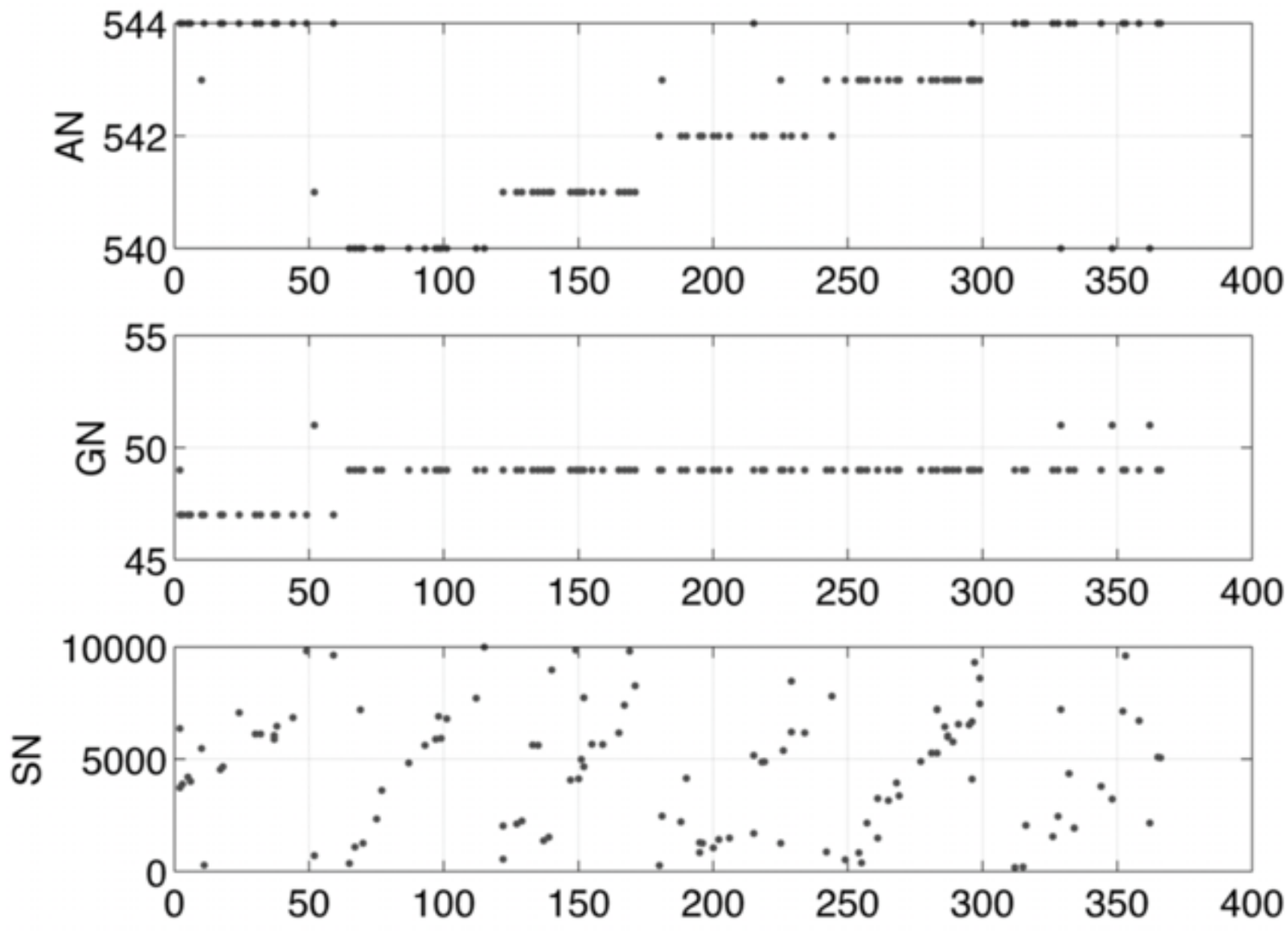
Information retrieval

- What are the resources we're comparing?
- How do we describe each one?
- How do we measure "similarity"
- Evaluation?

Modern IR

- Modern IR accounts for much more information than document similarity
    - Prominence/reliability of document (PageRank)
    - Geographic location
    - Search query history
- This can become a supervised problem to learn how to map these more elaborate features of a query/session to the search ranking. How do we represent our data?

Acquisti and Gross



SSN component assignment (y) as a function of date of birth (OR, 1996)