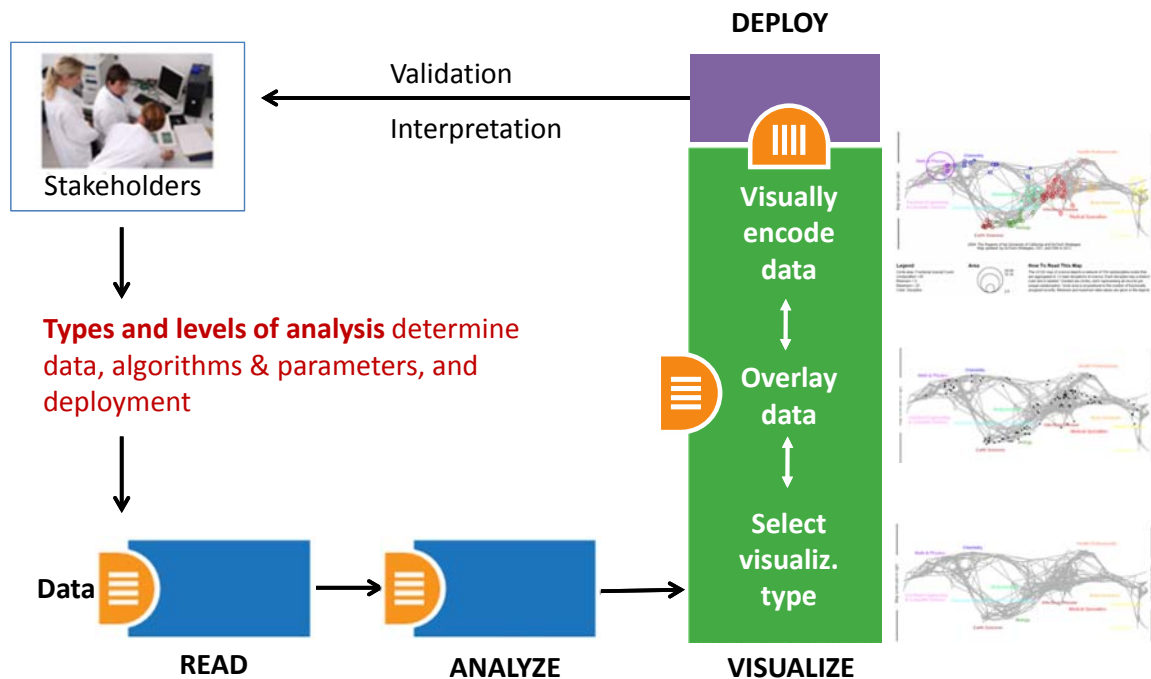


Needs-Driven Workflow Design



18

Information Visualization MOOC

Unit 4 – “What”: Topical Data

Workflow Design

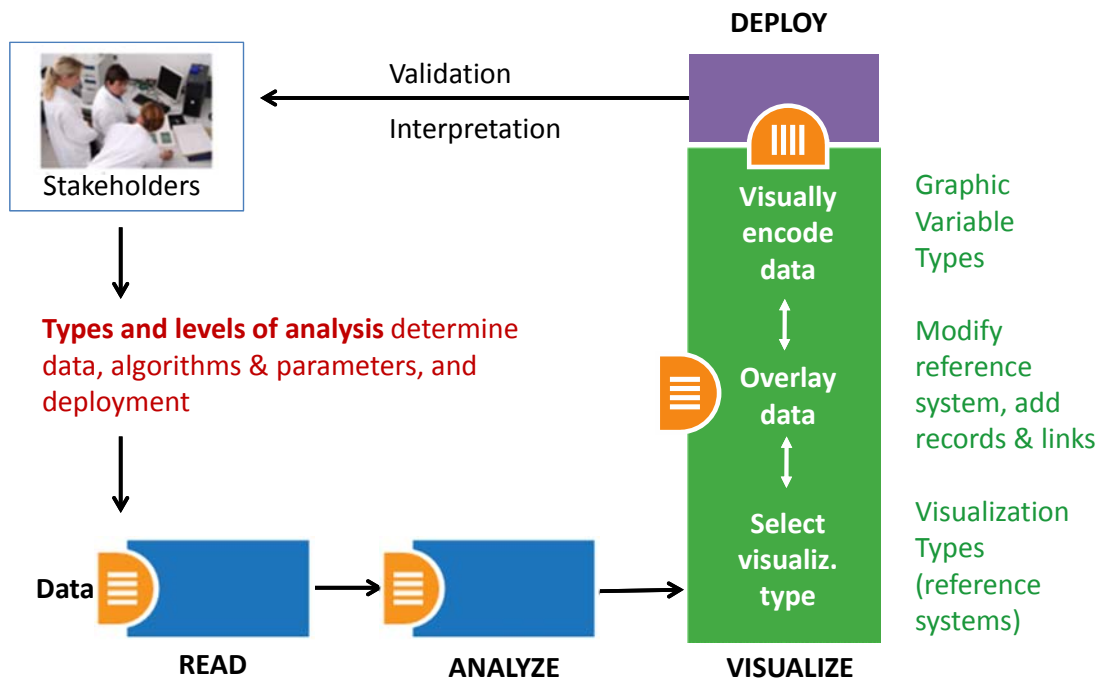
Relevant Research Disciplines:

Linguistics, Computer Science, Artificial Intelligence

Reference

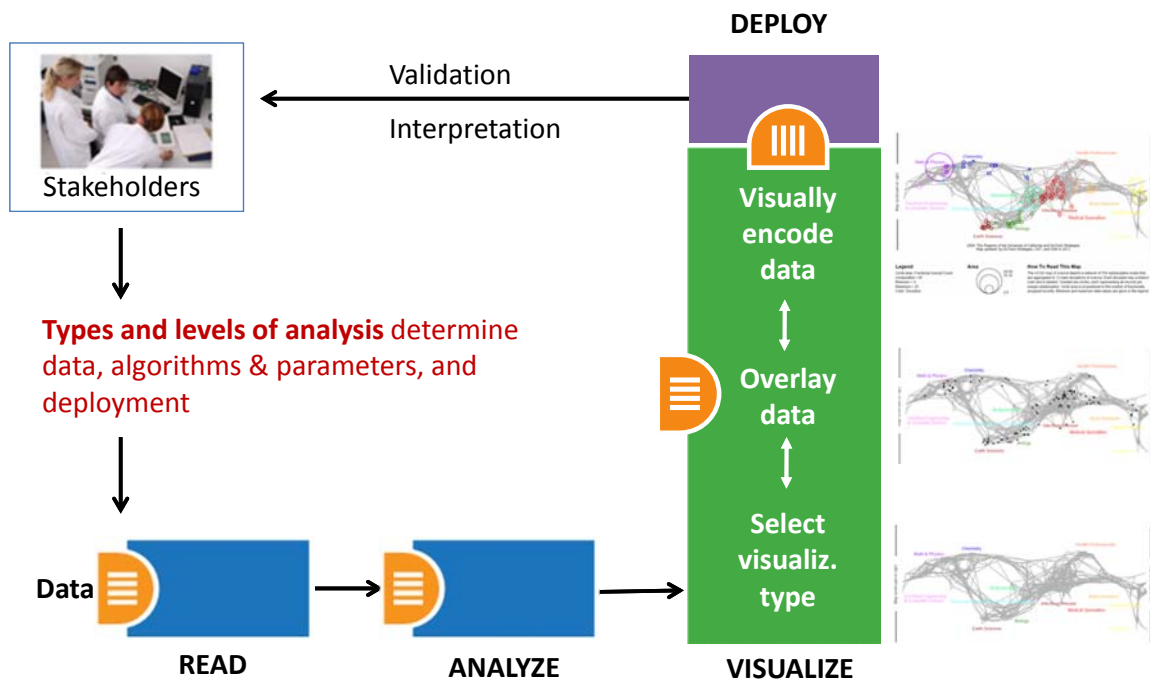
Börner, Katy, Chaomei Chen, and Kevin W. Boyack. 2003. [“Visualizing Knowledge Domains.”](#) Chap. 5 in *Annual Review of Information Science & Technology*, edited by Blaise Cronin, 37:179-255. Medford, NJ: American Society for Information Science and Technology.

Needs-Driven Workflow Design



20

Needs-Driven Workflow Design



21

Read Data

Data Repositories:

- KDD Datasets, <http://www.kdnuggets.com/datasets/>
- Digging into Data list of repositories, <http://www.diggingintodata.org/Repositories/tabid/167/Default.aspx>
- WordNet lexical database for English, <http://wordnet.princeton.edu/>
- [Google ngrams datasets](#), text from millions of books scanned by Google
- Scholarly Database, <http://sdb.cns.iu.edu>

Major Data Formats:

- TXT
- CSV

22

Preprocessing—Text Normalization

Sample text: Emergence of Scaling in Random Networks

- Lowercase: The example text becomes "emergence of scaling in random networks."
- Tokenize: The text blob is split into a list of individual words. The example text becomes "emergence|of|scaling|in|random|networks."
- Stem: Common or low-content prefixes and suffixes are removed to identify the core concept. The example text becomes "emerg|of|scale|in|random|network."
- Stopword: Low-content tokens like "of" and "in" are removed (see [the complete stopwords list](#)). The example text becomes "emerg|scale|random|network."
- Identification of synonymy and polysemy.

23

Topical Analysis

- Frequency analysis
- Clustering/Classification
- Sentiment analysis
- Burst analysis, see Unit 1
- Dimensionality reduction, see ARIST chapter.

24

Using a Dictionary and Thesaurus

Visual Thesaurus <http://www.visualthesaurus.com/vocabgrabber/>
Sorted by relevance, occurrences, select 'geography' words

The screenshot displays the Visual Thesaurus website interface. On the left, there are two panels showing word lists sorted by relevance. The top panel is titled 'Found 410 words' and shows a list of words including 'Man', 'life', 'ring', 'star', 'b', 'Lord', 'Night', 'one', 'rings', 'American', 'back', 'beautiful', 'b', 'episode', 'godfather', 'good', 'g', 'King', 'knight', 'love', 'man', 'men', 'once', 'part', 'princess', 'rain', 're', 'terminator', 'toy', 'train', 'two', 'w'. The bottom panel is also titled 'Found 410 words' and shows a list of words including 'Man', 'city', 'river', 'America', 'Arabia', 'Caribbean', 'Casablanca', 'country', 'district', 'Fargo', 'Green', 'Indiana', 'island', 'Lawrence', 'Leon', 'Manhattan', 'metropolis', 'North', 'Orange', 'reservoir', 'Rio', 'Rio Bravo', 'Rwanda', 'Torino', 'Virginia', 'Washington', 'waterfront'. On the right, there is a word cloud for the word 'Man', showing various related terms such as 'male', 'person', 'valet', 'gentleman', 'piece', 'human being', 'human', 'homo', 'adult male', 'serviceman', 'military personnel', 'military man', 'subordinate', 'foot soldier', 'underling', 'subsidiary', 'Isle of Man', 'valet de chambre', 'gentleman's', 'gentleman', 'river', 'piece', 'human being', 'human', 'homo', 'adult male', 'serviceman', 'military personnel', 'military man', 'subordinate', 'foot soldier', 'underling', 'subsidiary'.

25

Visualizing Topical Data

- **Charts:** Wordle Word cloud
- **Tables:** GRIDL
- **Graphs:** MDS plots, circular visualization, Crossmaps, Google n-gram
- **Geospatial maps:** SOM maps
- **Network graphs:** Tree visualizations, word co-occurrence networks, concept maps, science map overlays

Black ones are exemplified on subsequent slides.

26

Chart Example: Word Cloud

Wordle.net of Titles – create your own at <http://wordle.net>



Layout:	Oval space filling; frequent words are closer to center
Type font size:	Word frequency
Font color:	No meaning, but different colors help legibility

27

Graph these *case-sensitive* comma-separated phrases: Albert Einstein, Sherlock Holmes, Frankenstein

between 1800 and 2000 from the corpus English with smoothing of 3

Search lots of books

Share 1.3k

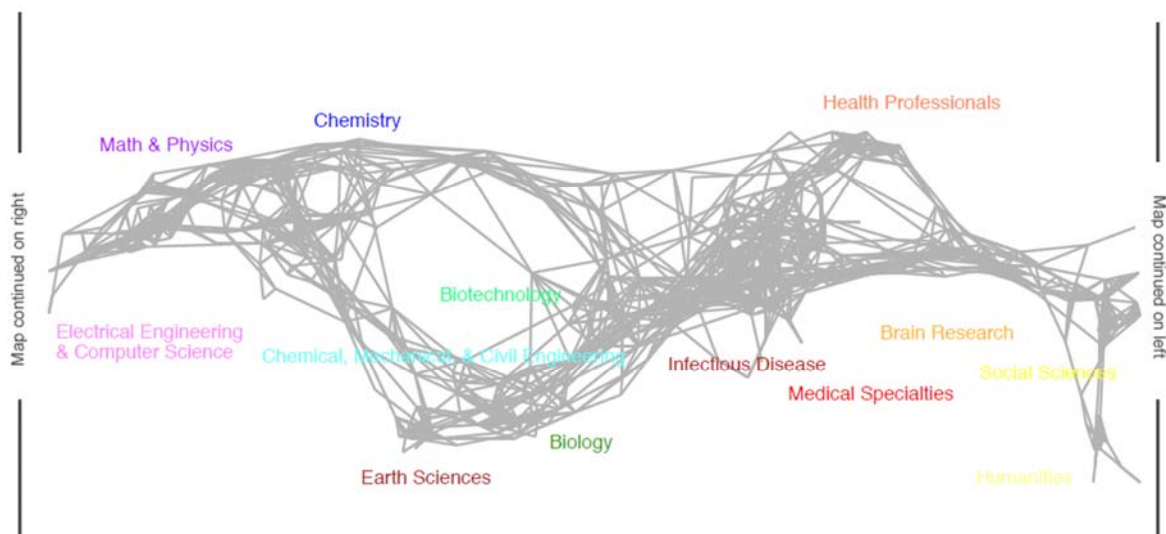
Tweet 1,176



An n-gram is a subsequence of n items from a given sequence. The items in question can be phonemes, syllables, letters, words, or base.

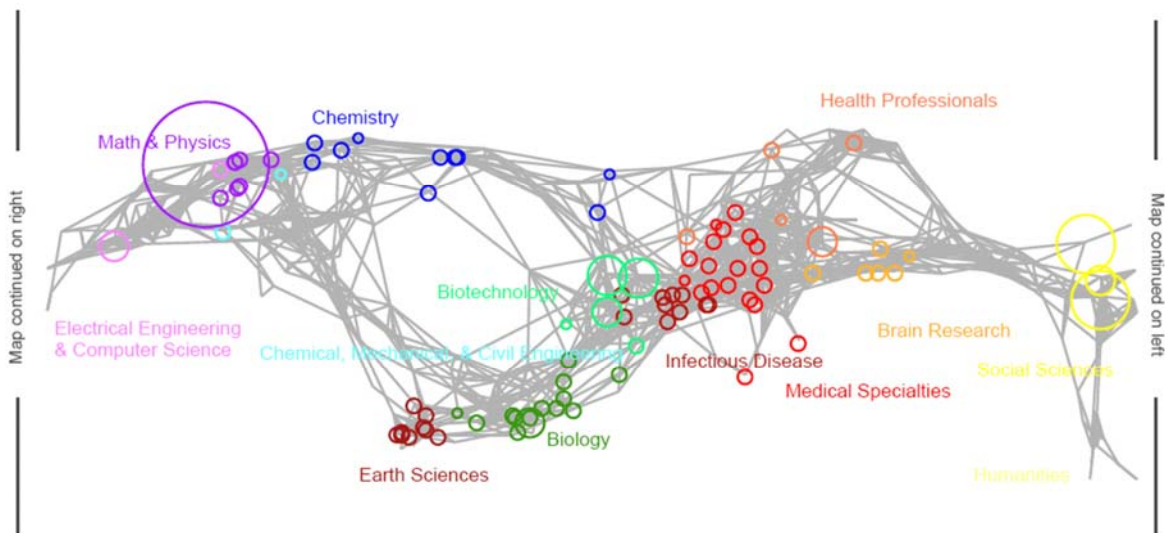
30

Network Graph-Science Map

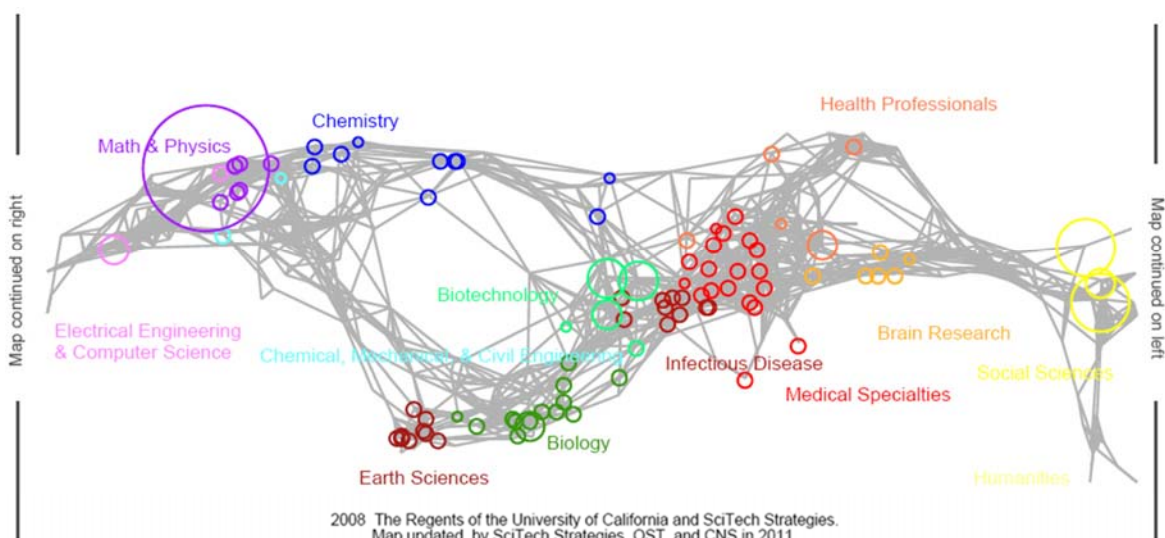


Reference system with proportional symbol overlay and legend. See also **Unit 4—Design and Update of a Classification System: The UCSD Map of Science**

31



32



2008 The Regents of the University of California and SciTech Strategies.
Map updated by SciTech Strategies, OST, and CNS in 2011.

Legend

Circle area: Fractional Journal Count
Unclassified = 95
Minimum = 0
Maximum = 25
Color: Discipline

Area



How To Read This Map

The UCSD map of science depicts a network of 554 subdiscipline nodes that are aggregated to 13 main disciplines of science. Each discipline has a distinct color and is labeled. Overlaid are circles, each representing all records per unique subdiscipline. Circle area is proportional to the number of fractionally assigned records. Minimum and maximum data values are given in the legend.

33

Co-Occurrence Network of IMDb Movie Title Words

Data retrieved from <http://www.imdb.com/chart/top> (on Nov 15, 2012).

Rank	Rating	Title	Votes
1.	9.2	The Shawshank Redemption (1994)	857,280
2.	9.2	The Godfather (1972)	625,241
3.	9.0	The Godfather: Part II (1974)	400,229
4.	8.9	Pulp Fiction (1994)	669,105
5.	8.9	The Good, the Bad and the Ugly	
6.	8.9	12 Angry Men (1957)	
7.	8.9	Schindler's List (1993)	
8.	8.9	The Dark Knight (2008)	

A	B	C	D	E
Rank	Rating	Title	Year	Votes
1	9.2	The Shawshank Redemption	1994	857,280
2	9.2	The Godfather	1972	625,241
3	9.0	The Godfather: Part II	1974	400,229
4	8.9	Pulp Fiction	1994	669,105
5	8.9	The Good, the Bad and the Ugly	1966	263,846
6	8.9	12 Angry Men	1957	211,144
7	8.9	Schindler's List	1993	444,891
8	8.9	The Dark Knight	2008	837,033

What words occur most frequently?
Which words are used together?

34

Text Preprocessing—Normalize

- Lowercase words
- Stem
- Remove stop words

Title		Title
The Shawshank Redemption		shawshank redempt
The Godfather		godfath
The Godfather: Part II		godfath ii
Pulp Fiction		pulp fiction
The Good, the Bad and the Ugly		good bad ugli
12 Angry Men		12 angri men
Schindler's List		schindler list
The Dark Knight		dark knight
The Lord of the Rings: The Return of the King		lord ring return king
Fight Club		fight club
Star Wars: Episode V - The Empire Strikes Back		star war episod v empir strike
One Flew Over the Cuckoo's Nest		flew cuckoo nest
The Lord of the Rings: The Fellowship of the Ring		lord ring fellowship ring
Inception		incept
Goodfellas		goodfella
Star Wars		star war
Seven Samurai		seven samurai
The Matrix		matrix

stopwords.txt
File Edit Format
a
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although

35

Text Preprocessing—Extract Network

- Tokenize
- Extract co-occurrence network

Without preprocessing:

Nodes: 470

Isolated nodes: 48

Edges: 905

With preprocessing:

Nodes: 405

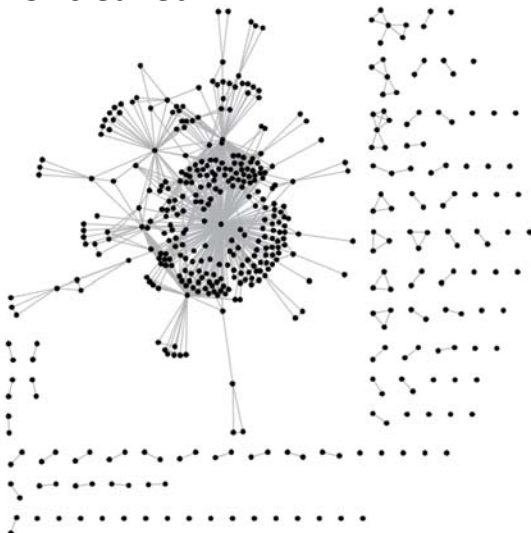
Isolated nodes: 75

Edges: 319

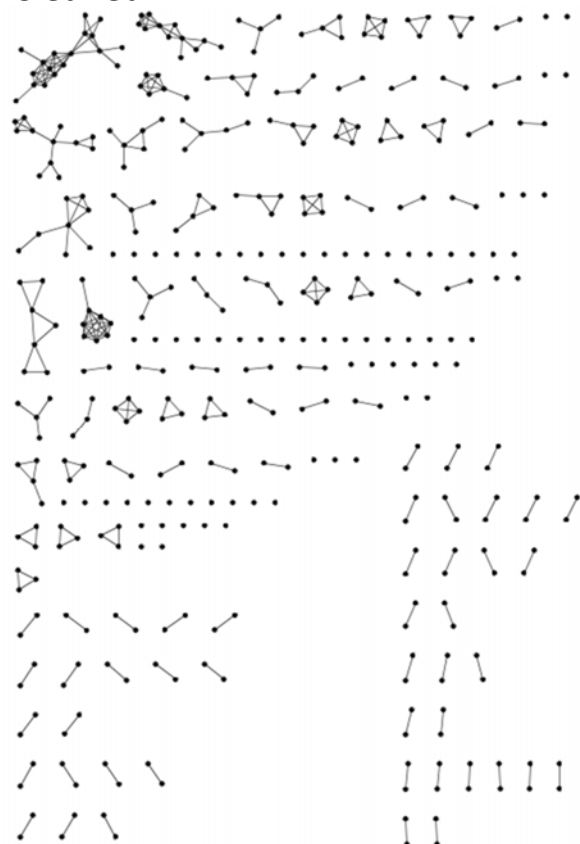
— A	
— About	Affair
Affairs	Afraid
Afraid	Afternoon
Afternoon	Alien
Alien	Amã©li
Aliens	Amadeus
— All	America
Amã©lie	American
Amadeus	Amor
America	Anatomi
American	Angri
Amores	Anni
Anatomy	Apart

36

Uncleaned

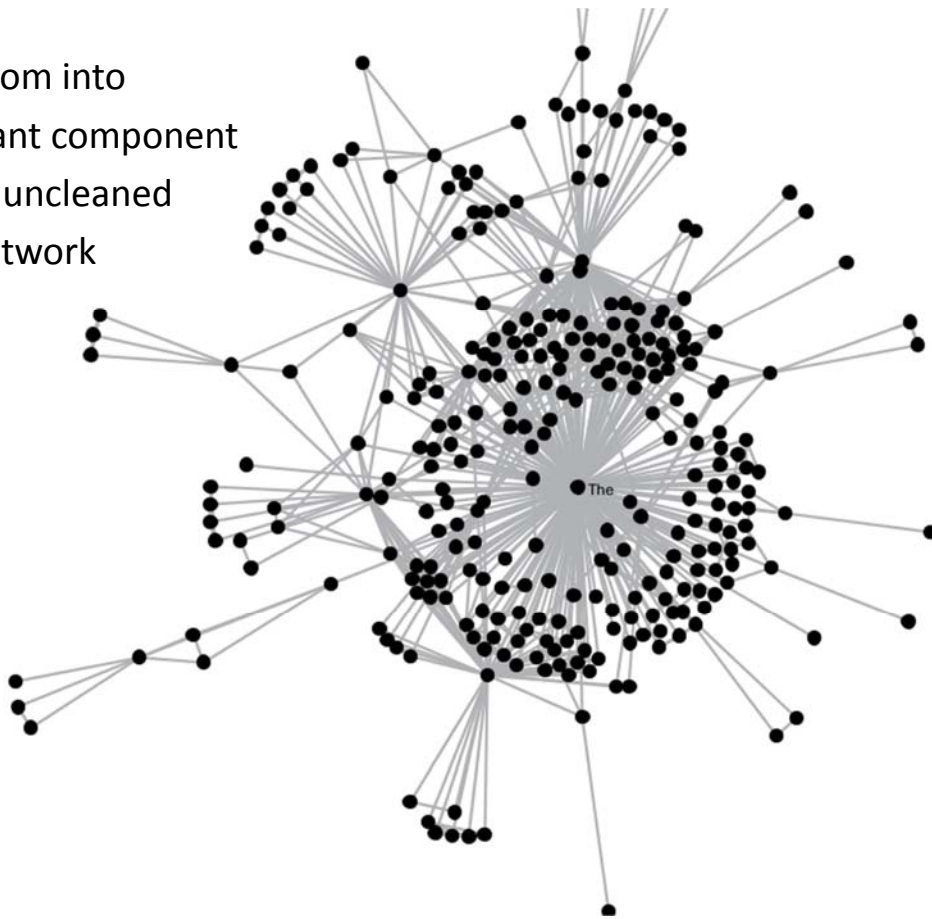


Cleaned



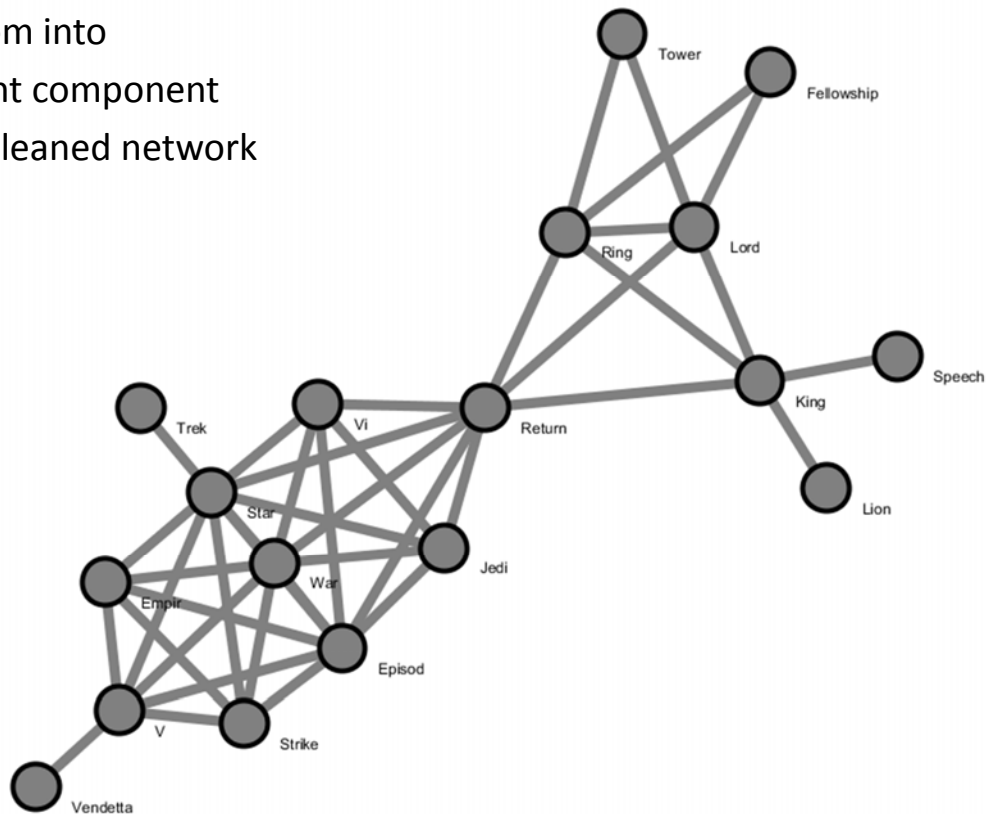
37

Zoom into
giant component
of uncleaned
network



38

Zoom into
giant component
of cleaned network



39

Movie Similarity Network Based on Shared Words

Four movies with 10 unique title words:

Title	Unique words										Total #words
	empir	episod	jedi	return	star	strike	trek	war	vi	v	
star war episod vi return jedi	0	1	1	1	1	0	0	1	1	0	6
star war episod v empir strike	1	1	0	0	1	1	0	1	0	1	6
star trek	0	0	0	0	1	0	1	0	0	0	2
star war	0	0	0	0	1	0	0	1	0	0	2

Semantic network of movies based on shared words:

*Vertices 4

1 "star war episod vi return jedi" 6

2 "star war episod v empir strike" 6

3 "star trek" 2

4 "star war" 2

*Edges 6

1 2 3

1 3 1

1 4 2

2 3 1

2 4 2

3 4 1

	star war episod vi return jedi	star war episod v empir strike	star trek	star war
1 star war episod vi return jedi		3	1	2
2 star war episod v empir strike	3		2	2
3 star trek	1	1		1
4 star war	2	2	1	

Complete matrix has 250 movies and 397 normalized words.

40

Movie Similarity Network

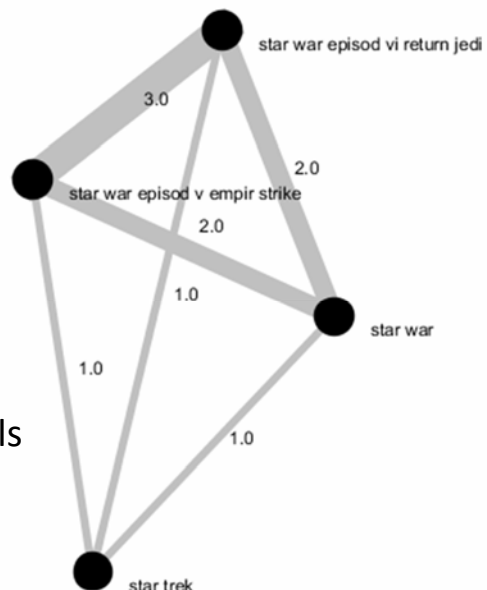
Semantic network of movies based on shared words:

But, titles have different length!

Solutions:

- Normalize by #words
- Use only important words (tf-idf)
- Apply LSA or topic detection
- Run Poisson-based language models

See also [Unit 4—Comparison of Text- and Linkage-Based Approaches](#)



Complete matrix has 250 movies and 397 normalized words.

41

Relevant Tools

- TextAnalyzer, <http://textalyser.net>
- TextTrend (OSGi/CIShell compatible), <http://textrend.org>
- VOSviewer, <http://vosviewer.com>

See many more at <http://www.kdnuggets.com/software/text.html>

Please post your favorite to Twitter, Flickr using tags “ivmooc” and “#topictools.”

