Text Visualization

Maneesh Agrawala

CS 294-10: Visualization
Spring 2011

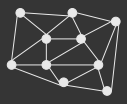
Last Time: Graph Layout 2

Graphs and Trees

Graphs

Model relations among data

Nodes and *edges*



Trees

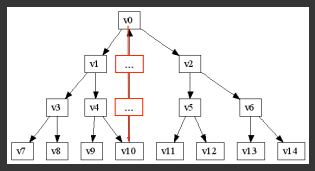
Graphs with hierarchical structure

- Connected graph with N-1 edges

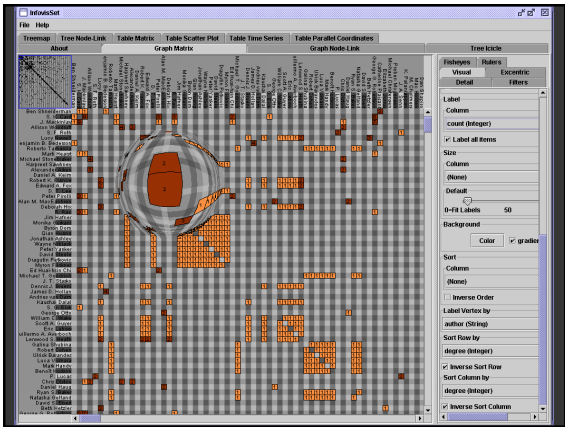
Nodes as *parents* and *children*

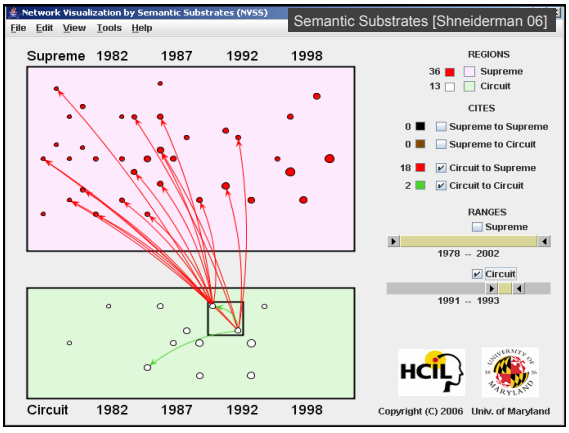
Sugiyama-style graph layout

Layer 1
Layer 2
Layer 3
Layer 4

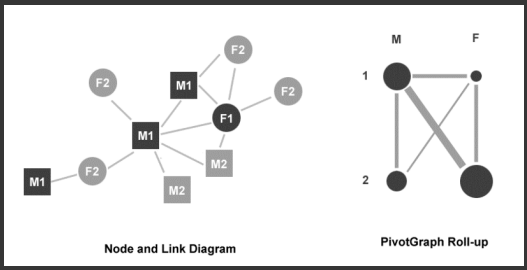


Reverse some edges to remove cycles
Assign nodes to hierarchy layers → Longest path layering
Create dummy nodes to "fill in" missing layers
Arrange nodes within layer, minimize edge crossings
Route edges – layout splines if needed



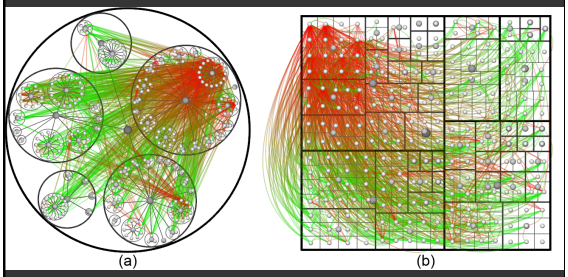


PivotGraph

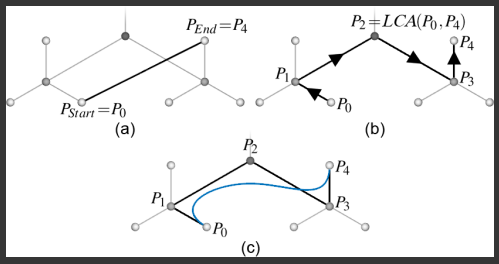


Hierarchical Edge Bundles

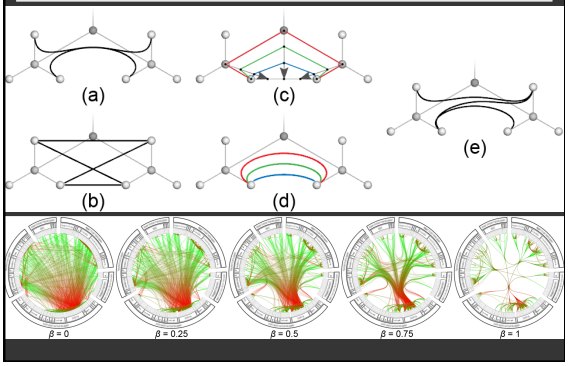
Trees with Adjacency Relations

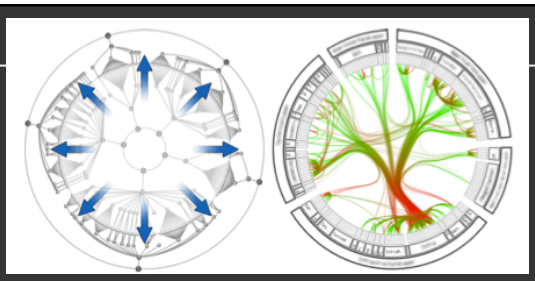


Bundle Edges along Hierarchy

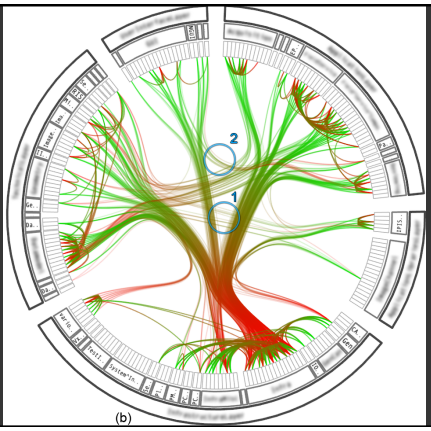


Configuring Edge Tension





Use radial tree layout for inner circle
Mirror to outside
Replace inner tree with hierarchical edge bundles



Summary

Tree Layout

- Indented / Node-Link / Enclosure / Layers
- How to address issues of scale?
  - Filtering and Focus + Context techniques

Graph Layout

- Tree layout over spanning tree
- Hierarchical “Sugiyama” Layout
- Optimization (Force-Directed Layout)
- Attribute-Driven Layout

Announcements

Final project

Design new visualization method

- Pose problem, Implement creative solution

Deliverables

- Implementation of solution
- 8-12 page paper in format of conference paper submission
- 2 design discussion presentations

Schedule

- Project proposal: 3/14
- Project presentation: 4/4
- Final paper and presentation: 5/3 1:30-3pm 6th floor Soda

Grading

- Groups of up to 3 people, graded individually
- Clearly report responsibilities of each member

Text Visualization

Why visualize text?

Why Visualize Text?

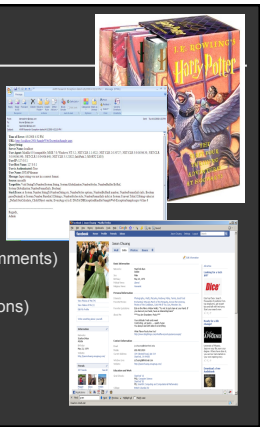
Understanding: get the "gist" of a document

Grouping: cluster for overview or classification

Compare: compare document collections, or inspect evolution of collection over time

Correlate: compare patterns in text to those in other data, e.g., correlate with social network

What is text data?



Documents

- Articles, books and novels
- Computer programs
- E-mails, web pages, blogs
- Tags, comments

Collection of documents

- Messages (e-mail, blogs, tags, comments)
- Social networks (personal profiles)
- Academic collaborations (publications)

Challenge: Visualize Dissertations

You have 20 years of university Ph.D. theses:

- Text
- Year
- Dept.
- Author
- Advisor
- Committee

What questions might you want to answer?

What visualizations might help?

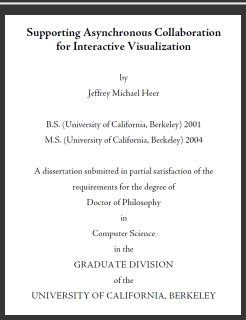
A Concrete Example

What would help you gauge?

Topics in document?

Relationship to other docs?

…



Supporting Asynchronous Collaboration for Interactive Visualization

by

Jeffrey Michael Heer

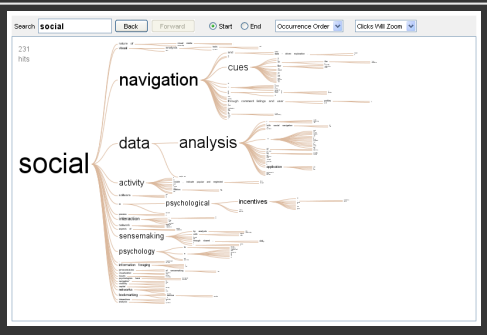
B.S. (University of California, Berkeley) 2001

M.S. (University of California, Berkeley) 2004

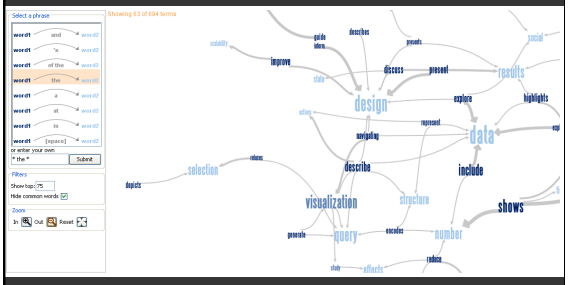
A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science in the GRADUATE DIVISION of the UNIVERSITY OF CALIFORNIA, BERKELEY

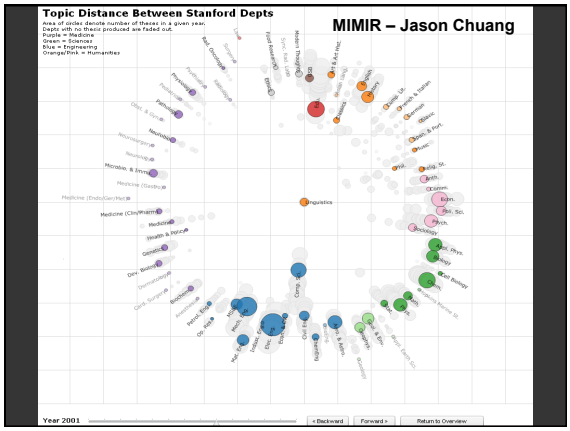
Tag Cloud: Word Counts

Word Tree: Word Sequences



PhraseNet: “A the B”





A Double Gulf of Evaluation

Many (most?) text visualizations do not represent text directly, they represent a model

- term statistics
- clusters
- …

Can you interpret the visualization?

- How well does it convey the properties of the model?

Do you trust the model?

- How does the model enable us to reason about the text?

Lessons for Text Visualization

Show (or provide access to) source text

- Let readers assess model
- Let readers use visualization as index into documents

Find meaningful abstractions for grouping docs

- Are clusters interpretable?

Where possible use text to represent text… but which terms are the most descriptive?

Topics

Text as Data
Visualizing Document Content
Evolving Documents
Visualizing Conversation
Document Collections

Text as Data

Words are (not) nominal?

High dimensional (10,000+)

More than equality tests

Words have meanings and relations

- Correlations: Hong Kong, San Francisco, Bay Area
- Order: April, February, January, June, March, May
- Membership: Tennis, Running, Swimming, Hiking, Piano
- Hierarchy, antonyms & synonyms, entities, …

Text Processing Pipeline

Tokenization: segment text into *terms*

Special cases? e.g., "San Francisco", "L'ensemble", "U.S.A."

Remove stop words? e.g., "a", "an", "the", "to", "be"?

Stemming: one means of normalizing terms

Reduce terms to their "root"; Porter's algorithm for English

e.g., *automate(s), automatic, automation* all map to *automat*

For visualization, want to reverse stemming for labels

- Simple solution: map from stem to the most frequent word

Result: ordered stream of terms

The Bag of Words Model

Ignore ordering relationships within the text

A document ~ vector of term weights

Each term corresponds to a dimension (10,000+)

Each value represents the relevance

- For example, simple term counts

Aggregate into a document x term matrix

Document vector space model

Document x Term matrix

Each document is a vector of term weights

Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

WordCount (Harris 2004)

http://wordcount.org

Weaknesses of Tag Clouds

Sub-optimal visual encoding (size vs. position)

Inaccurate size encoding (long words are bigger)

May not facilitate comparison (unstable layout)

Term frequency may not be meaningful

Does not show the structure of the text

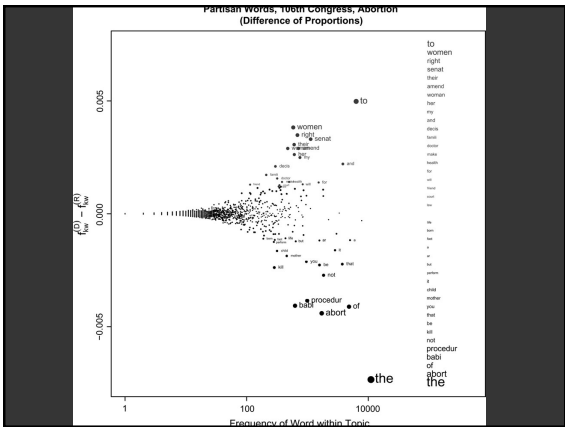
Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

Can take log frequency: $\log(1 + tf_{td})$

Can normalize to show proportion: $tf_{td} / \Sigma_t tf_{td}$



Keyword Weighting

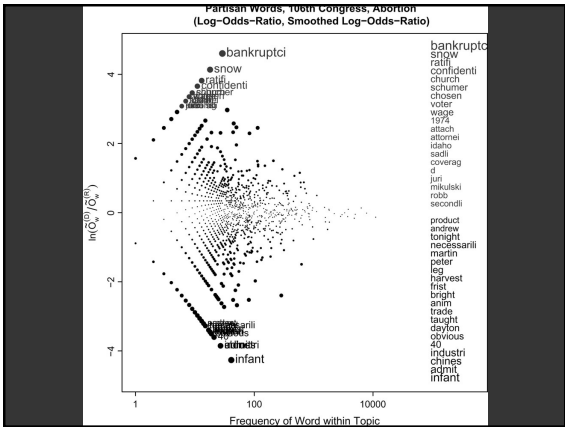
Term Frequency

tf_{td} = count(t) in d

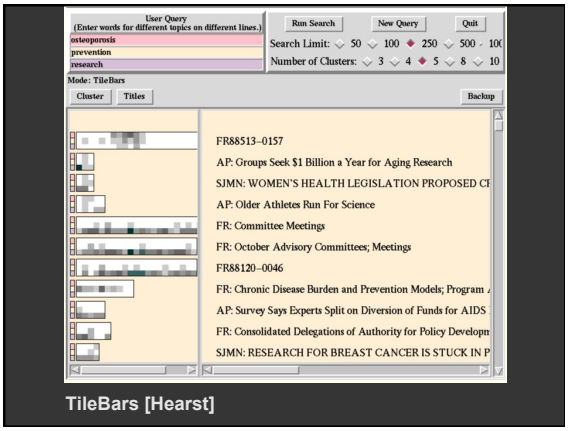
TF.IDF: Term Freq by Inverse Document Freq

$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$

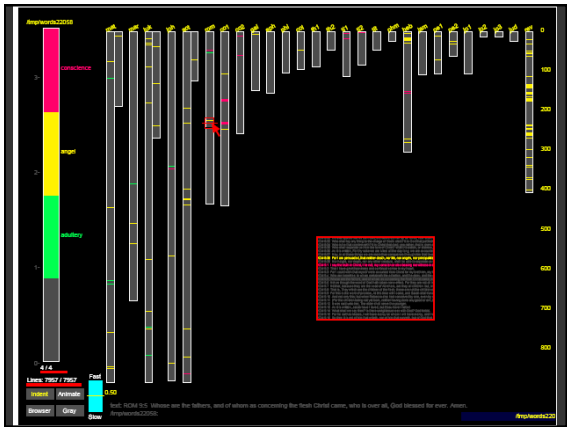
df_t = # docs containing t; N = # of docs

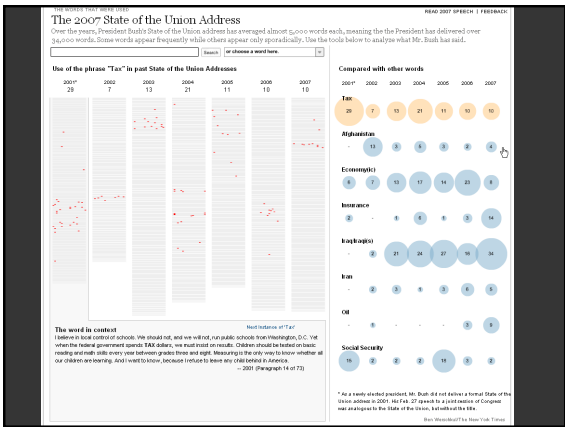


Visualizing Document Content

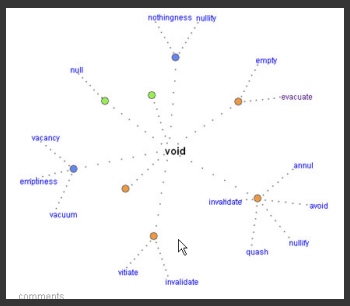


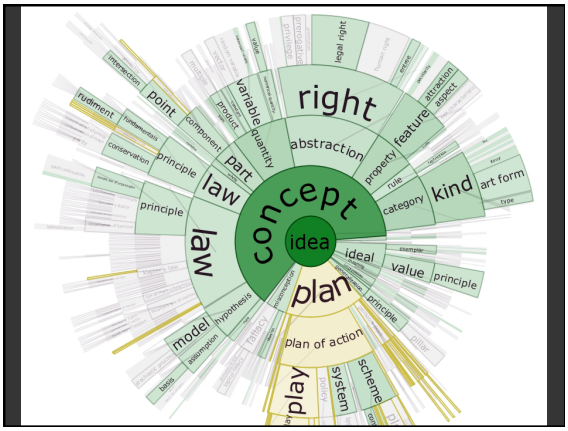
TileBars [Hearst]





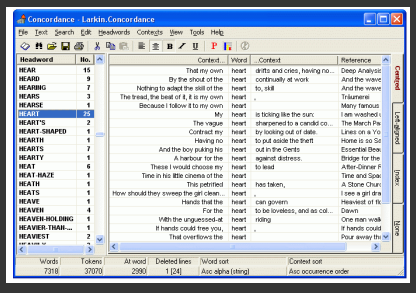
Visual Thesaurus [ThinkMap]

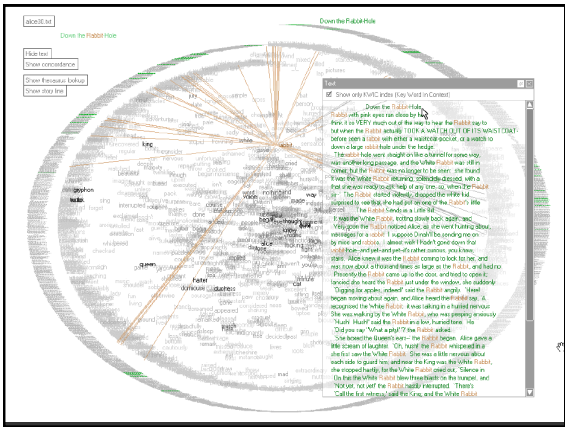


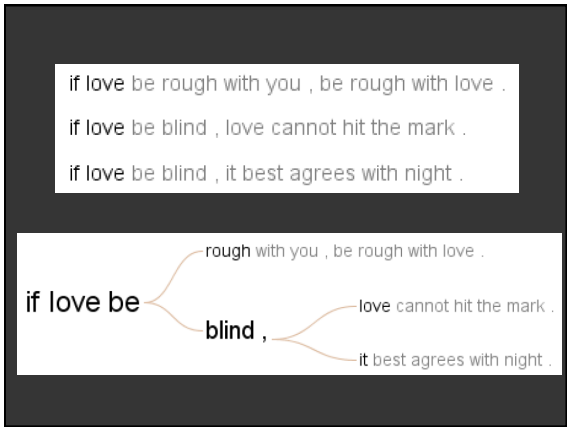


Concordance

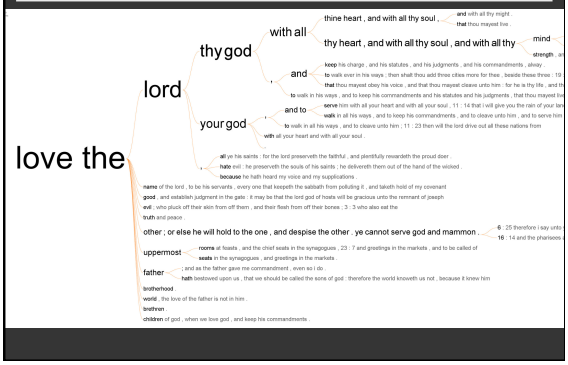
What is the common local context of a term?



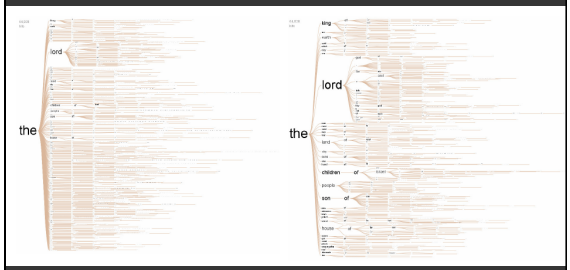




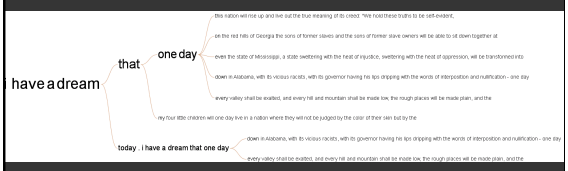
WordTree (Wattenberg et al)

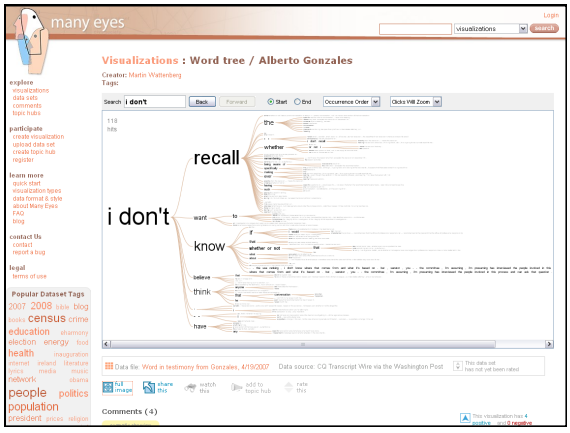


Filter infrequent runs



Recurrent themes in speech





Glimpses of structure

Concordances show local, repeated structure
But what about other types of patterns?

For example

Lexical: <A> at <B>

Syntactic: <Noun> <Verb> <Object>

Phrase Nets [van Ham et al]

Look for specific linking patterns in the text:

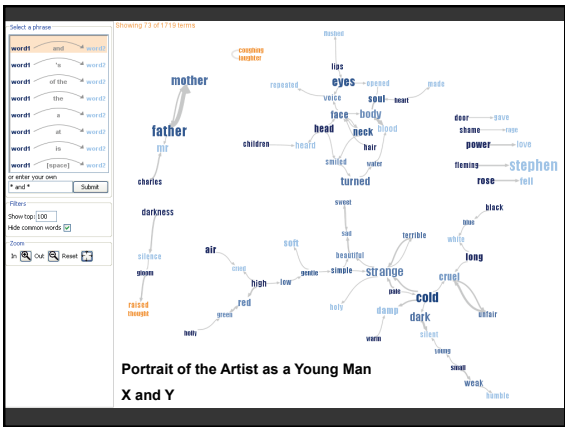
'A and B', 'A at B', 'A of B', etc

Could be output of regexp or parser

Visualize extracted patterns in a node-link view

Occurrences → Node size

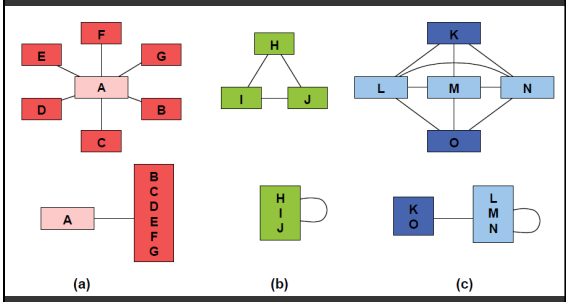
Pattern position → Edge direction

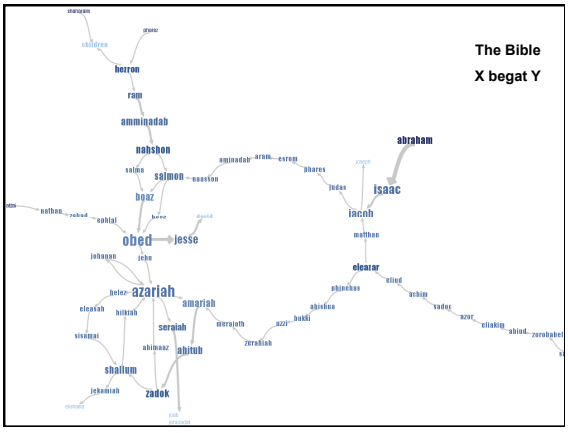


Portrait of the Artist as a Young Man

X and Y

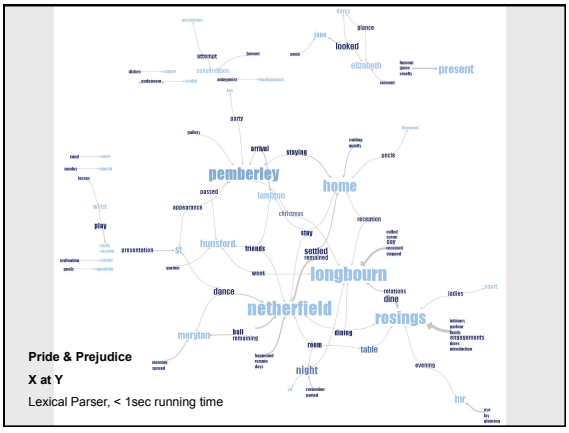
Node Grouping





The Bible

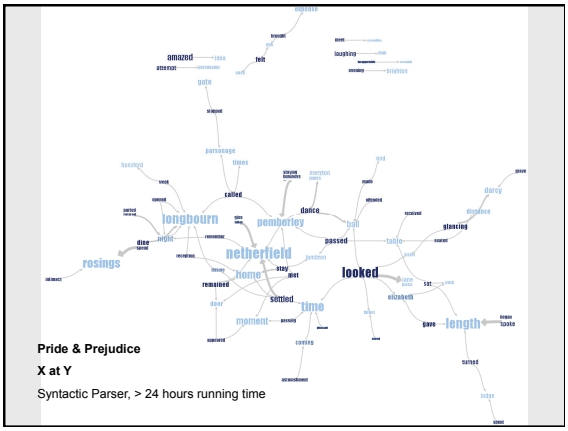
X begat Y

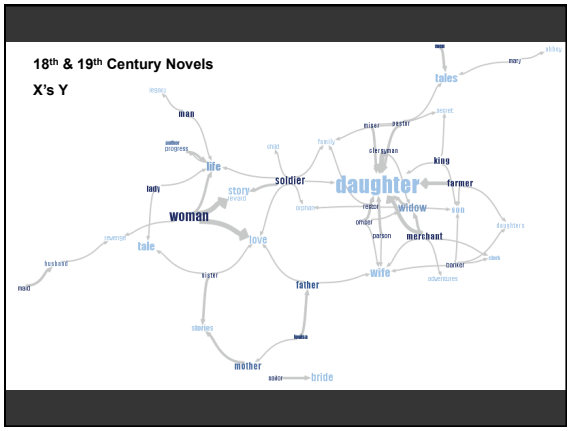


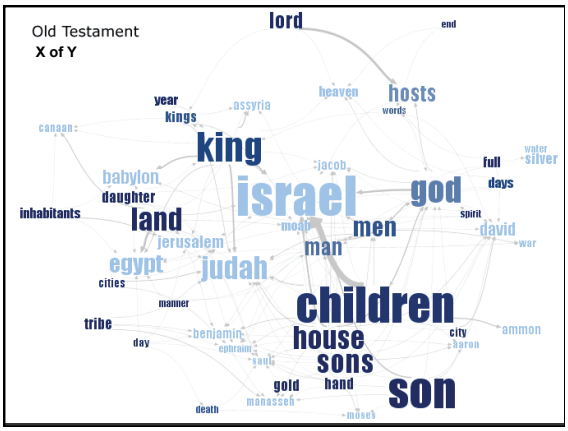
Pride & Prejudice

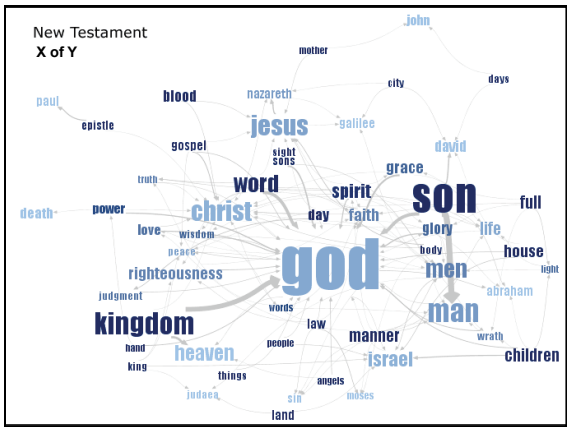
X at Y

Lexical Parser, < 1sec running time

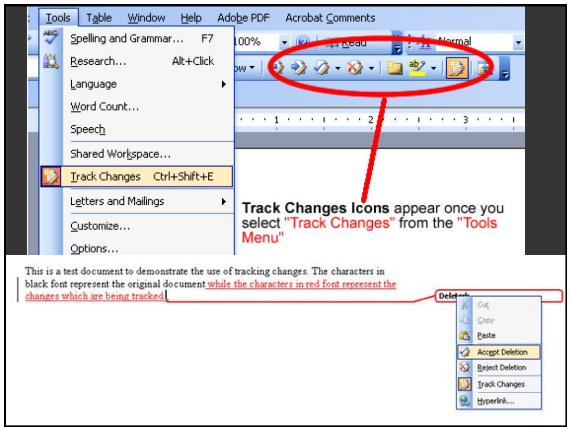








Evolving Documents



Visualizing Revision History

How to depict contributions over time?

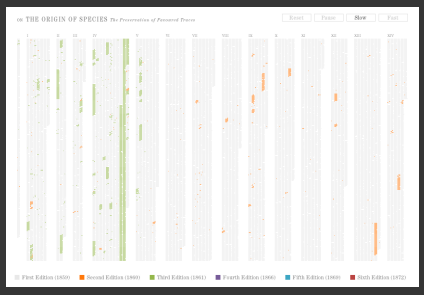
Example: Wikipedia history log

Chocolate
Revision history

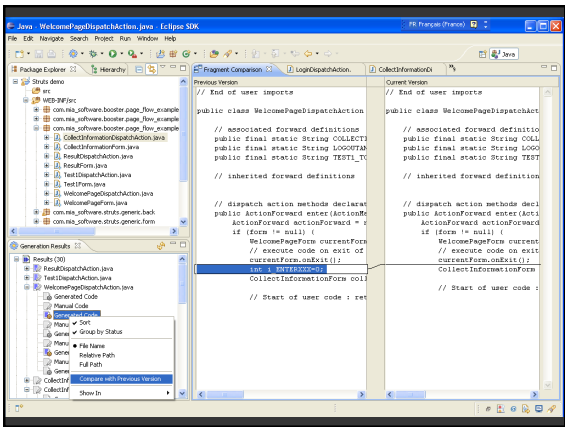
Legend: (cur) = difference with current version, (last) = difference with preceding version, M = minor edit

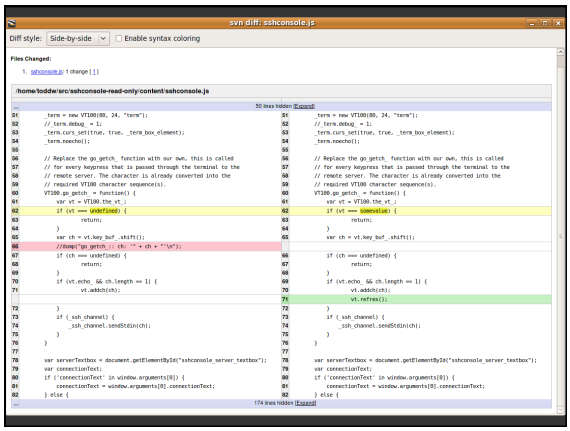
- (cur) (last) . . 12:01, 20 Aug 2003 . . Dysprosia *(neaten to do, rearrange see also)*
- (cur) (last) . . 11:59, 20 Aug 2003 . . Patrick
- (cur) (last) . . 11:52, 20 Aug 2003 . . 81.203.98.109
- (cur) (last) . . **M** 18:36, 6 Aug 2003 . . Maniac *(corrected spelling)*
- (cur) (last) . . 18:32, 6 Aug 2003 . . Daniel Quinlan *(removing obscure heraldry information, belongs on [[heraldry]] if anywhere)*
- (cur) (last) . . 15:21, 6 Aug 2003 . . Rnhermen
- (cur) (last) . . 15:08, 6 Aug 2003 . . Cyp *(Chocolate often has odd shapes.)*
- (cur) (last) . . 19:14, 3 Aug 2003 . . Daniel C. Boyer *("chocolate" as shade of gules in heraldry)*
- (cur) (last) . . **M** 02:00, 30 Jul 2003 . . Evercat *(fmt)*

Animated Traces (Ben Fry)

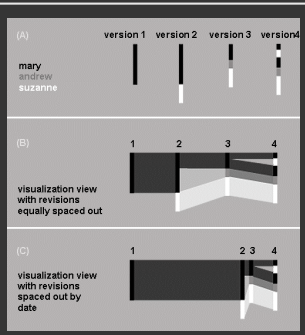


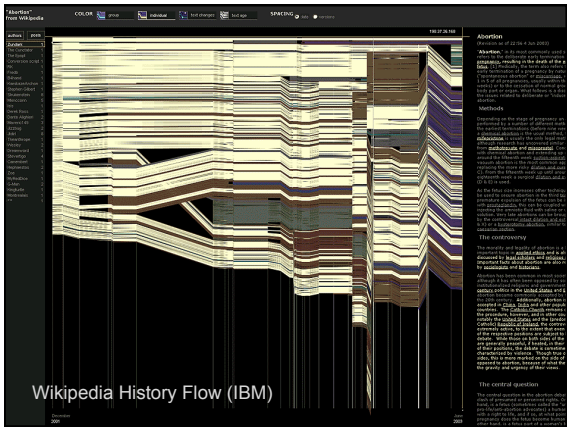
http://benfry.com/traces/



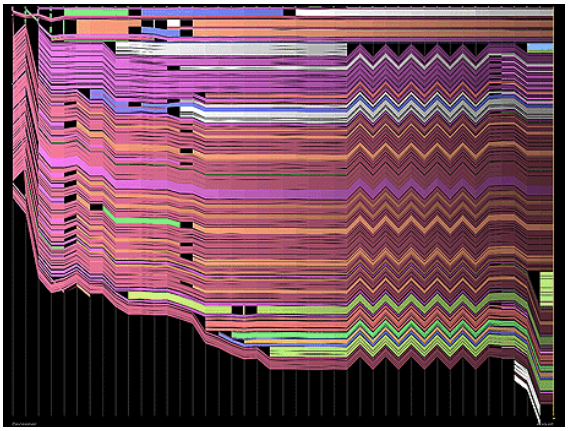


History Flow (Viégas et al)





Wikipedia History Flow (IBM)



Visualizing Conversation

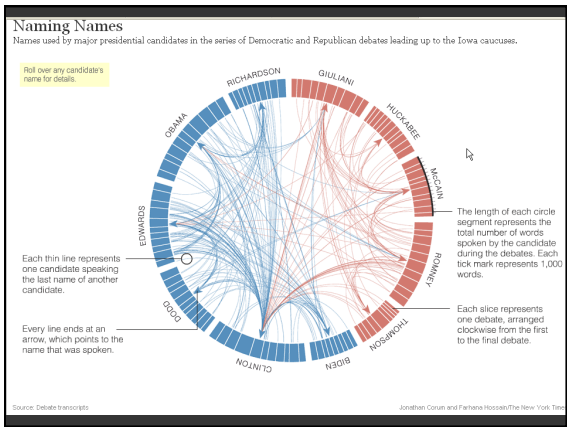
Visualizing Conversation

Many dimensions to consider:

- Who (senders, receivers)
- What (the content of communication)
- When (temporal patterns)

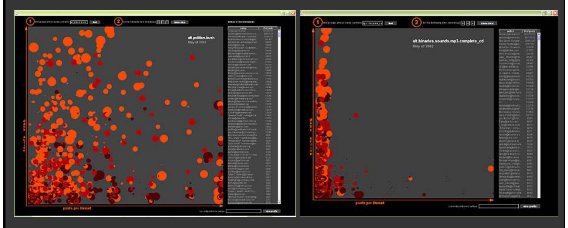
Interesting cross-products:

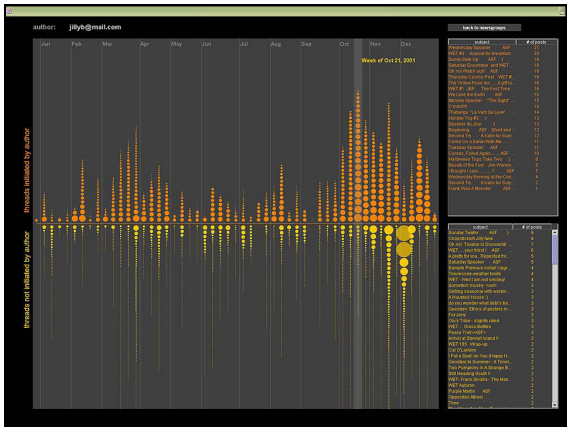
- What x When → Topic "Zeitgeist"
- Who x Who → Social network
- Who x Who x What x When → Information flow



Usenet Visualization (Viégas & Smith)

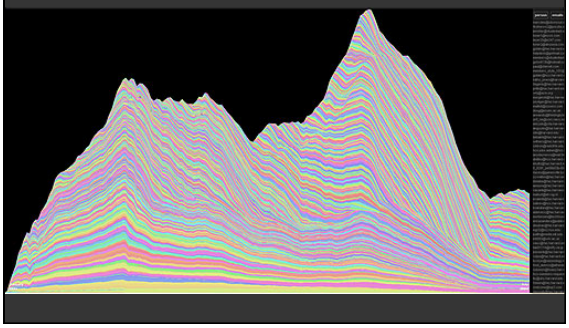
Show correspondence patterns in text forums
Initiate vs. reply; size and duration of discussion



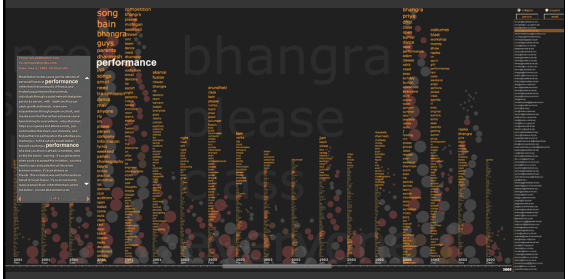


Mountain (Viégas)

Conversation by person over time (who x when)



Themail (Viégas et al)



One person over time, TF.IDF weighted terms