Information Visualization

Introduction

Petra Isenberg (+ Jean-Daniel Fekete)
petra.isenberg@inria.fr (jean-daniel.fekete@inria.fr)

After today you will…

- have gained an overview of the research area
- learned basic principles of data representation and interaction

Why

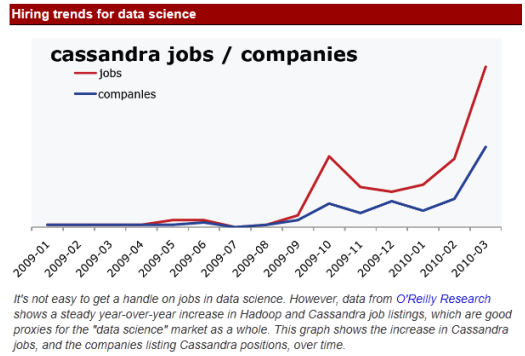
INFORMATION VISUALIZATION

It is estimated that 800 exabyte (800x 10^{19}) of digital information will be generated this year

4

[source: The Diverse and Exploding Digital Universe, IDC, 2008]
[credit: Did You Know; Fisch, McLeod, Brenman]

5



It's not easy to get a handle on jobs in data science. However, data from O'Reilly Research shows a steady year-over-year increase in Hadoop and Cassandra job listings, which are good proxies for the "data science" market as a whole. This graph shows the increase in Cassandra jobs, and the companies listing Cassandra positions, over time.

"The ability to take data -- to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades."

Hal Varian, chief economist at Google

Question

how can we effectively access data?

- understand its structure?
- make comparisons?
- make decisions?
- gain new knowledge?
- convince others?
- ...

7

Many possible ways to address…

Example

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Raw Data from Anscombe's Quartet

[Source: Anscombe's quartet, Wikipedia]

Statistical Analysis

For all four columns, the statistics are identical

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

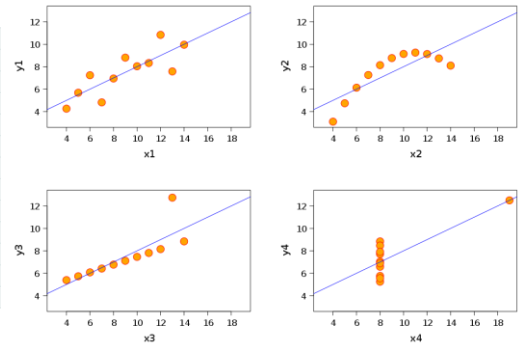
Mean of x	9.0
Variance of x	11.0
Mean of y	7.5
Variance of y	4.12
Correlation between x and y	0.816
Linear regression line	$y = 3 + 0.5x$

[Source: Anscombe's quartet, Wikipedia]

Visual Representation of the Data

Visual representation reveals a different story

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



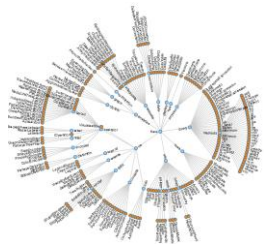
11

[Source: Anscombe's quartet, Wikipedia]

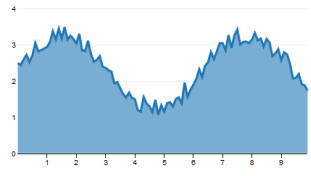
Why visual data representations?

- Vision is our most dominant sense
- We are very good at recognizing visual patterns
- We need to see and understand in order to explain, reason, and make decisions

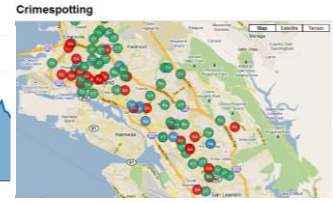
common examples:



graphs / hierarchies



charts

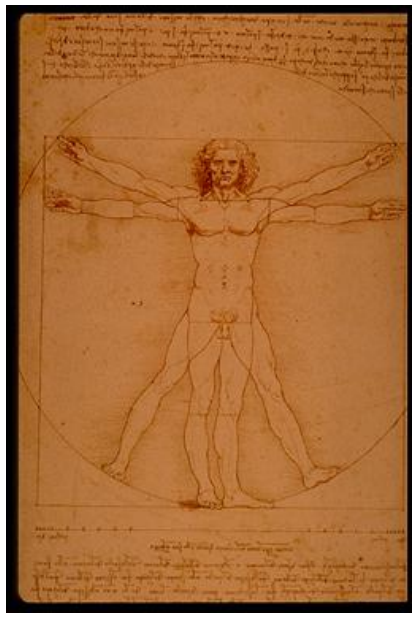


maps

all examples from: http://vis.stanford.edu/protovis/

Other benefits of visualization

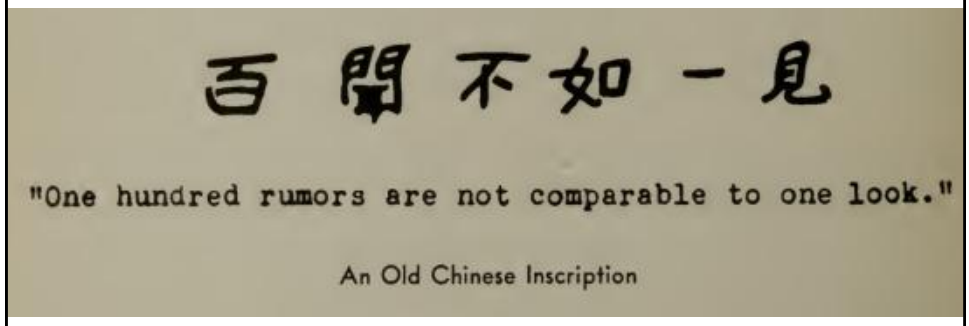
- expand human working memory
  - offload cognitive resources to the visual system,
- reduce search
  - by representing a large amount of data in a small space,
- enhance the recognition of patterns
  - by making them visually explicit
- aid monitoring of a large number of potential events
- provides a manipulable medium & allows exploration of a space of parameter values.



L'occhio,
che si dice finestra dell'anima,
è la principale via donde il comune
senso può piú copiosamente e
magnificamente considerare
le infinite opere di natura.

Leonardo da Vinci
(1452 - 1519)

The eye…
the window of the soul,
is the principal means
by which the central sense
can most completely and
abundantly appreciate
the infinite works of nature.



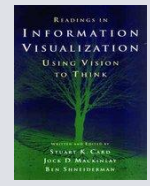
Via Brinton, Graphic Presentation, 1939

Information visualization

- Create visual representation
- Concentrates on abstract data
- Includes interaction

Official Definition:

The use of computer-supported, interactive, visual representations of abstract data to amplify cognition.
[Card et al., 1999]



Functions of Visualizations

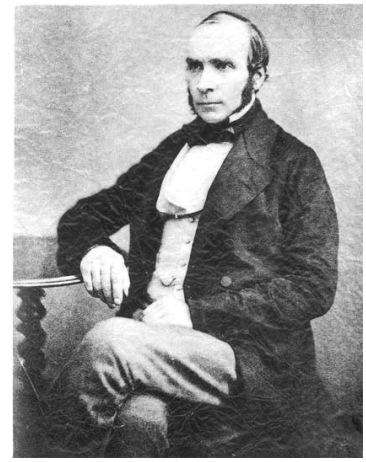
- Recording information
  - Tables, blueprints, satellite images
- Processing information
  - needs feedback and interaction
- Presenting information
  - share, collaborate, revise
  - for oneself, for one's peers and to teach
- Seeing the unseen

Visualization of abstract data has been practiced for hundreds of years…

HISTORICAL EXAMPLES

The Broadway Street Pump

- In 1854 cholera broke out in London
  - 127 people near Broad Street died within 3 days
  - 616 people died within 30 days
- "Miasma in the atmosphere"
- Dr. John Snow was the first to link contaminated water to the outbreak of cholera
- How did he do it?
  - he talked to local residents
  - identified a water pump as a likely source
  - used maps to illustrate his theory
  - convinced authorities to disable the pump



More info here: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

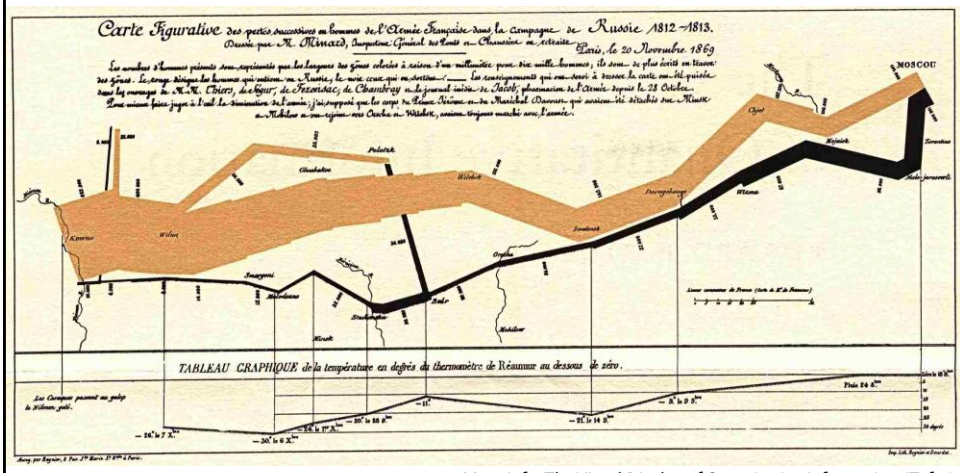
20

John Snow, 1854

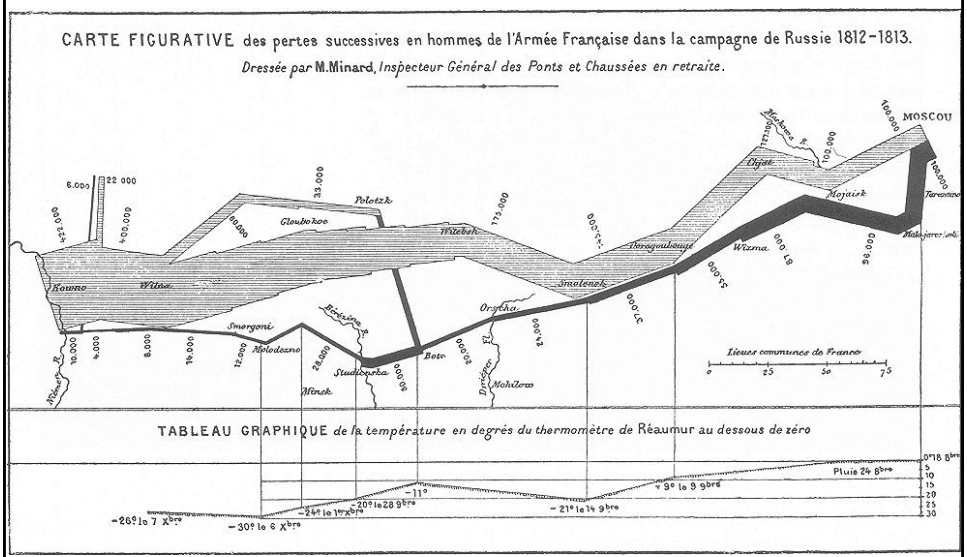
Napoleon's March on Moscow Charles Minard, 1869

Named the best statistical graphic ever drawn (by Edward Tufte)

- Includes: spatial layout linked with stats on: army size, temperature, time
- Tells a story in one overview



More info: The Visual Display of Quantitative Information (Tufte)

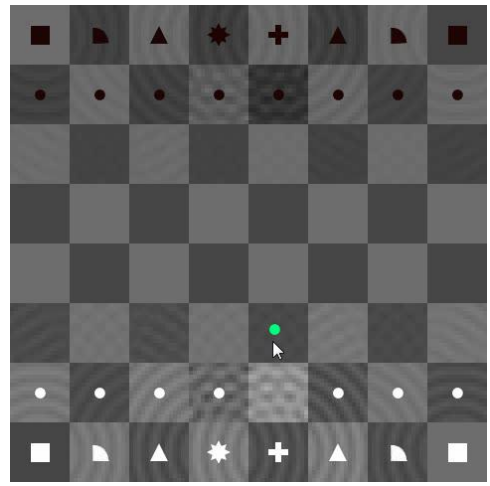


… AND VERY RECENTLY

TrashTrack

Winner of the NSF International Science & Engineering Visualization Challenge!
http://senseable.mit.edu/trashtrack/

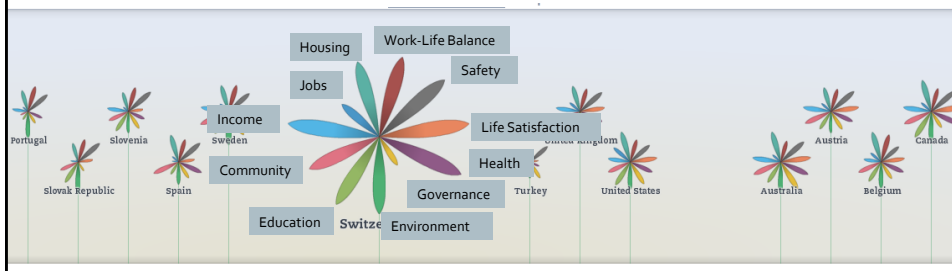
Artificial Intelligence



http://www.turbulence.org/spotlight/thinking/chess.html

Open Data

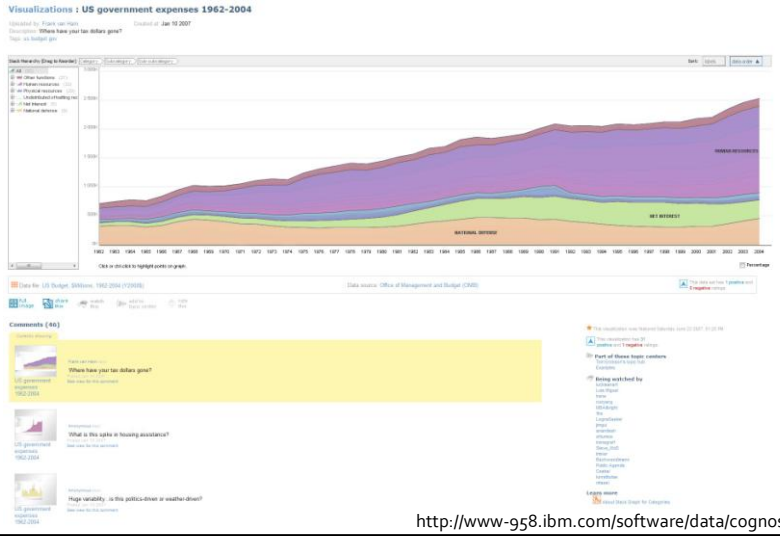
- Movement making government data freely available
- Encourage participation by everyone



OECD Better Life Index: http://www.oecdbetterlifeindex.org/

Many Eyes

- Upload data, create visualizations, discuss
- Distributed asynchronous collaboration

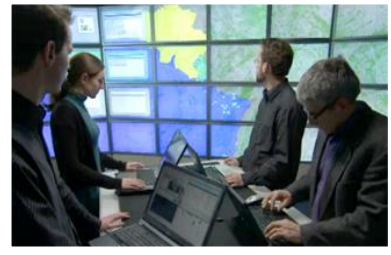


http://www-958.ibm.com/software/data/cognos/manyeyes/

Specific Visualization Environments

Molecular visualisation in the Reality Cube
University of Groningen, NL

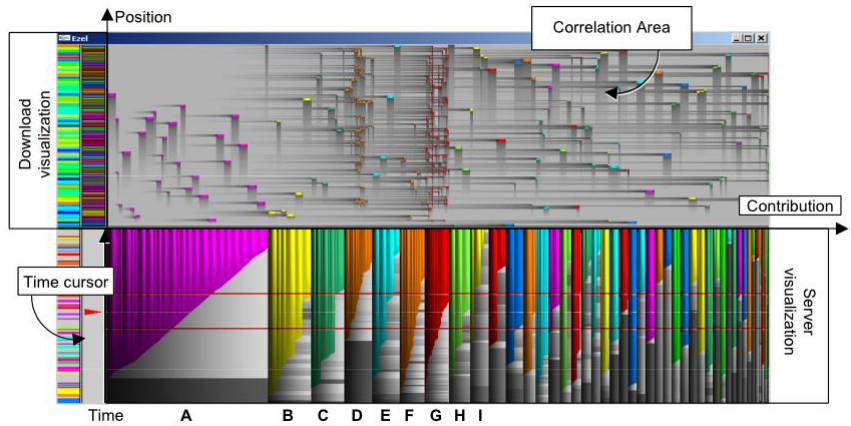
Tabletops for Visualization
University of Calgary



WILD Wall, INRIA

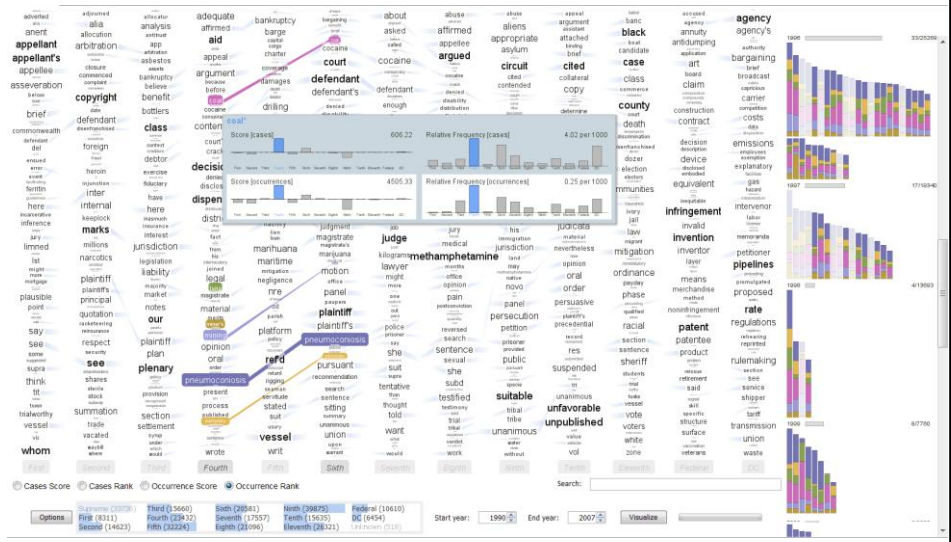
Software Visualization

EZEL: a Visual Tool for Performance Assessment of Peer-to-Peer File-Sharing Networks (Voinea et al., InfoVis, 2004)



Text Visualization

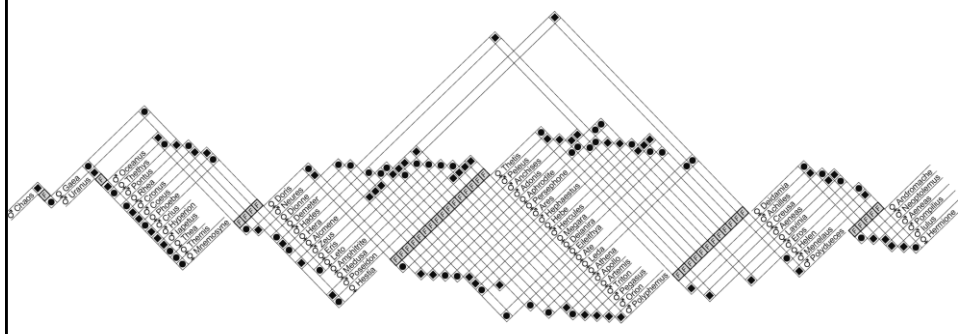
Parallel Tag Clouds to Explore Faceted Text Corpora (Collins et al., VAST 2009)



Graphs

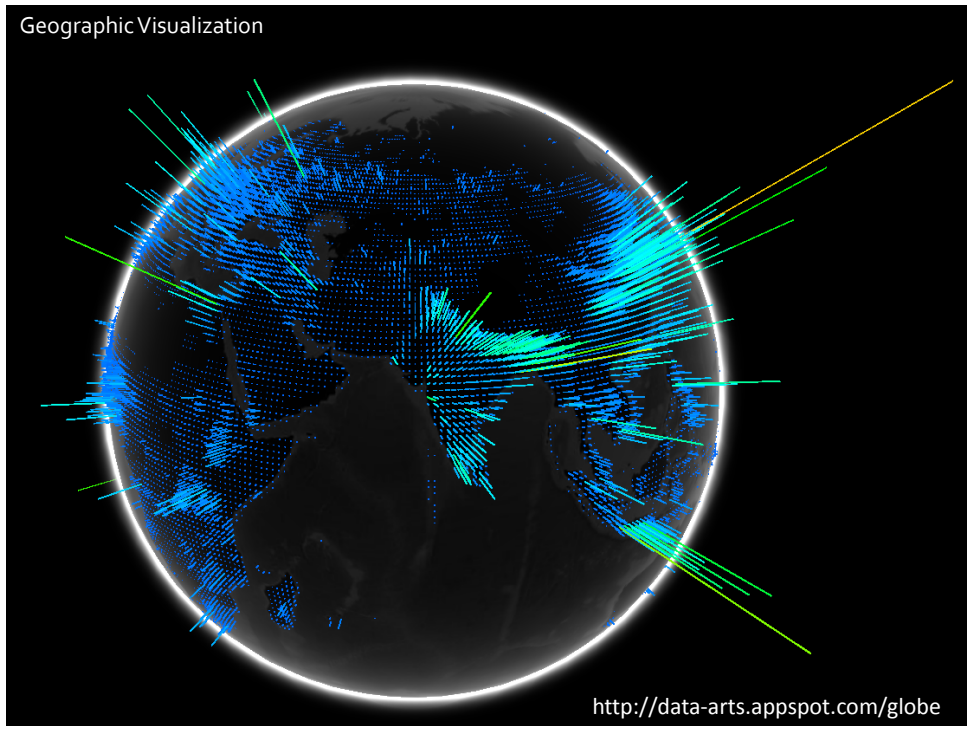
http://www.facebook.com/note.php?note_id=469716398919
Visualizing Friendships by Paul Butler on Tuesday, December 14, 2010

Family Trees



http://www.aviz.fr/geneaquilts/

Geographic Visualization



http://data-arts.appspot.com/globe

Weather

http://weatherspark.com/

Data Dashboards

http://globalspirometry.com

Resources for more examples

- Visualization conferences
- Blogs
  - http://infosthetics.com/
  - http://fellinlovewithdata.com/
  - http://eagereyes.org/
  - http://flowingdata.com/
  - http://www.informationisbeautiful.net/
- Books
  - Textbooks
    - Readings in Information Visualization: Using Vision to Think (a bit old now but good intro)
    - Information Visualization (Robert Spence – a light intro, I recommend as a start)
    - Information Visualization Perception for Design (Colin Ware, focused on perception and cognition)
    - Interactive Data Visualization: Foundations, Techniques, and Applications (Ward et al. – most recent)
  - Examples
    - Beautiful Data (McCandless)
    - Now You See it (Few)
    - Tufte Books: Visual Display of Quantitative Information (and others)
    - … (many more, ask me for details)

It is difficult to create

CREATE VISUALIZATIONS

GOOD

What is a representation?

- A representation is
  - a formal system or mapping by which the information can be specified (D. Marr)
  - a sign system in that it stands for something other than its self.
- for example: the number thirty-four

34 decimal

100010 binary

XXXIV roman

Presentation

- different representations reveal different aspects of the information

  decimal: counting & information about powers of 10,

  binary: counting & information about powers of 2,

  roman: impress your friends (outperformed by positional system)

- presentation

  how the representation is placed or organized on the screen

  34, **34**, 34

Principles of Graphical Excellence

- Well-designed presentation of interesting data – a matter of *substance*, *statistics*, *design*
- Complex ideas communicated with clarity, precision, efficiency
- Gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
- Involves almost always multiple variables
- Tell the truth about the data

The Visual Display of Quantitative Information, Tufte

41

Or a bit more simply…

- Solving a problem simply means representing it so as to make the solution transparent … (Simon, 1981)
- Good representations:
  - allow people to find relevant information
    - information may be present but hard to find
  - allow people to compute desired conclusions
    - computations may be difficult or "for free" depending on representations

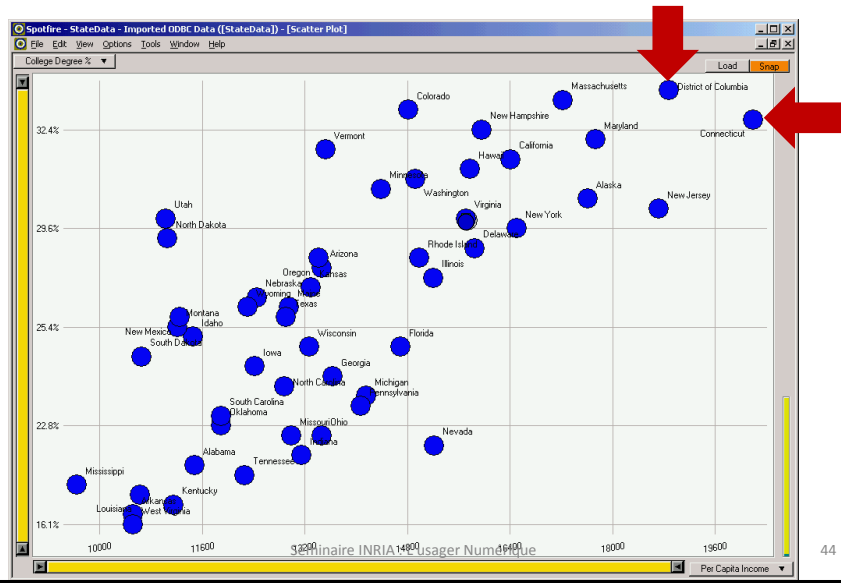
Good representation?

Table - StateData ()

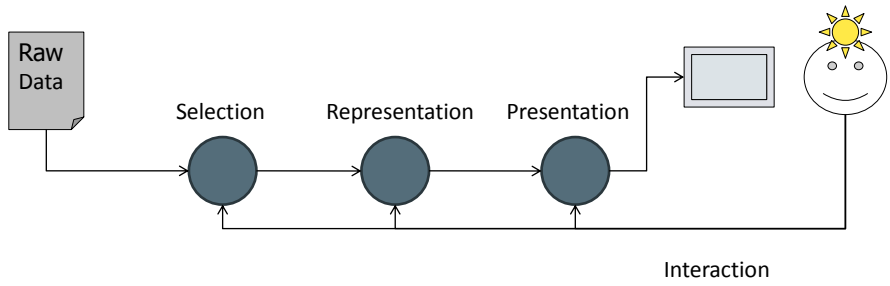
Load | Snap

State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20189
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14698
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13527
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

Good representation!



How do we arrive at a visualization?

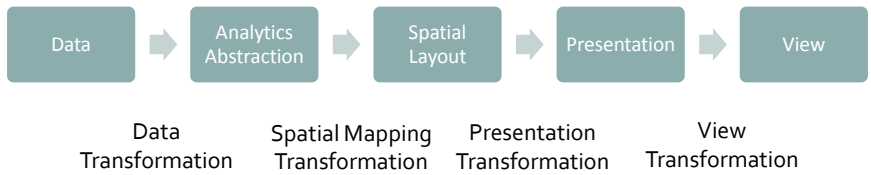


The Visualization Pipeline

From [Spence, 2000]

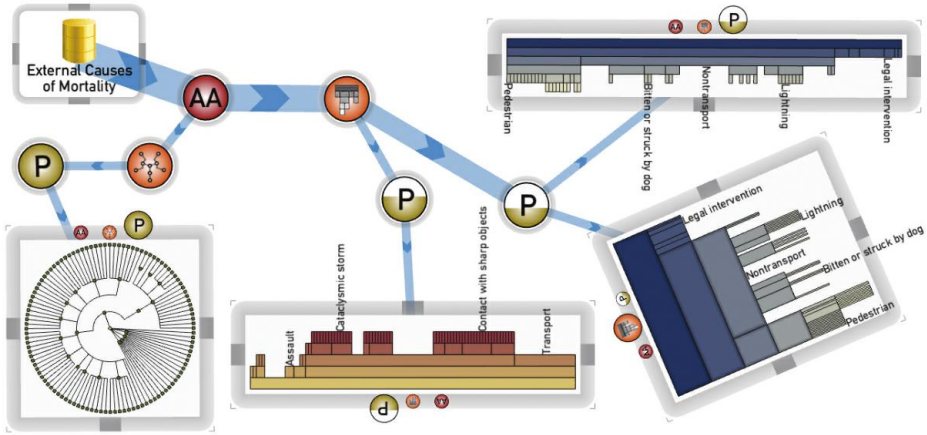
Visualization Reference Model

Also a visualization pipeline a bit expanded



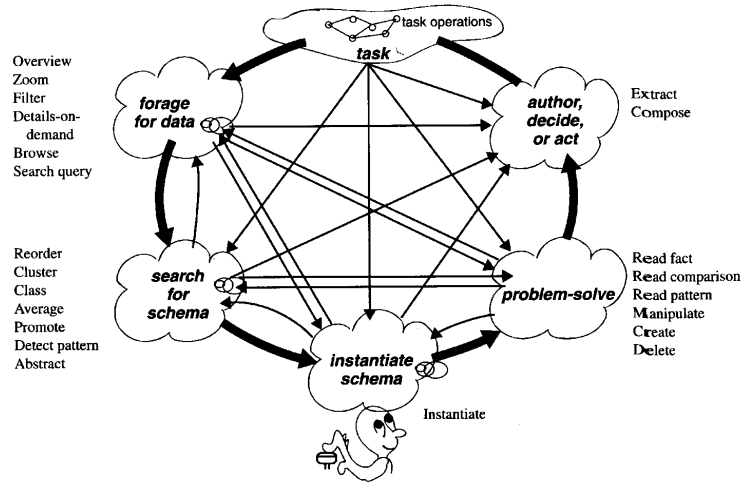
From [Card et al., Readings in Information Visualization]

Visualization pipeline in an image



[Tobiasz et al., 2009]

Knowledge Crystallization Cycle



Working with visualizations in NOT a linear process

[Card et al., 1999]

Pitfalls

- Selecting the wrong data
- Selecting the wrong data structure
- Filtering out important data
- Failed understanding of the types of things that need to be shown
- Choosing the wrong representation
- Choosing the wrong presentation format
- Inappropriate interactions provided to explore the data

Recap

- So far you
  - learned what information visualization is
  - learned about the advantages of visualization
  - saw a number of examples (historical and new)
  - tried to create your own first visualization from a dataset
- Next
  - you will get to know your data
  - you will learn about the basic components of visualization
  - try another example

Data

- Data is the foundation of any visualization
- The visualization designer needs to understand
  - the data properties
  - know what meta-data is available
  - know what people want from the data

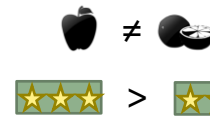
Nominal, Ordinal and Quantitative

- Nominal (labels)
  - Fruits: apples, oranges
- Ordered
  - Quality of meat: grade A, AA, AAA
  - Can be counted and ordered, but not measured
- Quantitative: Interval
  - no clear zero (or arbitrary)
  - e.g. dates, longitude, latitude
  - usually compare differences (intervals)
- Quantitative: Ratio
  - meaningful origin (zero)
  - physical measurements (temperature, mass, length)
  - counts and amounts

S.S. Stevens, On the theory of scales of measurements, 1946

Nominal, Ordinal and Quantitative

- Nominal (labels)
  - Operations: =, ≠
- Ordered
  - Operations: =, ≠, <, >
- Quantitative: Interval
  - Operations: =, ≠, <, >, -, +
  - Can measure distances or spans
- Quantitative: Ratio
  - Operationrs: =, ≠, <, >, - , +, ×, ÷
  - Can measure ratios or proportions



[1989 – 1999] + [2002 – 2012]

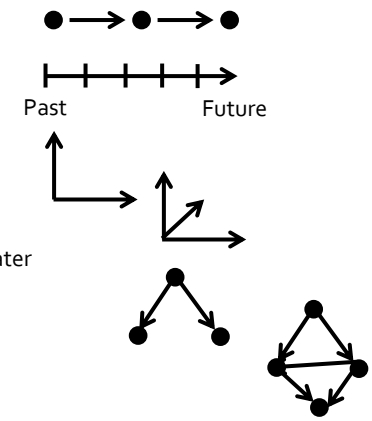
10kg / 5kg

S.S. Stevens, On the theory of scales of measurements, 1946

Data-Type Taxonomy

- 1D (linear)
- Temporal
- 2D (maps)
- 3D
- nD (relational)
- Trees (hierarchies)
- Networks (graphs)

vis examples later



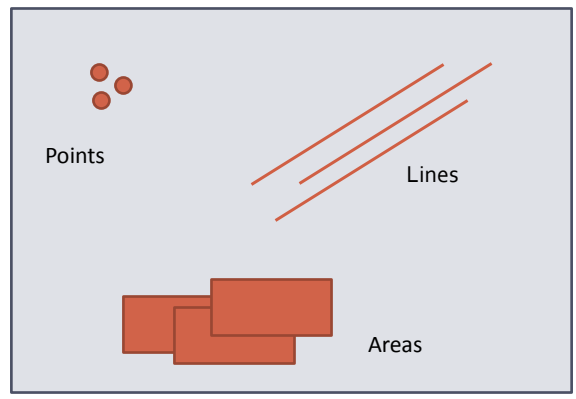
Shneiderman: The Eyes Have It

Why is this important?

- Nominal, ordinal, and quantitative data are best expressed in different ways visually
- Data types often have inherent tasks
  - temporal data (comparison of events)
  - trees (understand parent-child relationships)
  - …
- But:
  - any data type (1D, 2D,…) can be expressed in a multitude of ways!

Visualization's Main Building Blocks

Marks which represent:

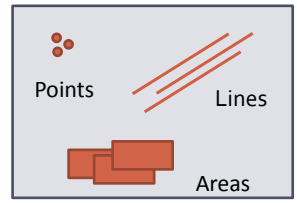


From Semiology of Graphics (Bertin)

The following slides on the topic adapted from Sheelagh Carpendale

Points

- "A point represents a location on the plane that has **no theoretical length or area**. This signification is independent of the size and character of the mark which renders it visible."
- a location
- marks that indicate points can vary in all visual variables

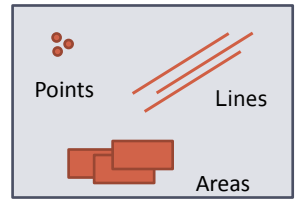


From Semiology of Graphics (Bertin)

63

Lines

- "A line signifies a phenomenon on the plane which has **measurable length but no area**. This signification is independent of the width and characteristics of the mark which renders it visible."
- a boundary, a route, a connection

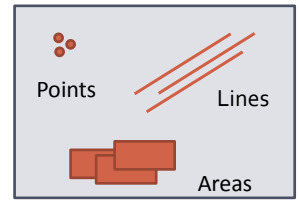


From Semiology of Graphics (Bertin)

64

Areas

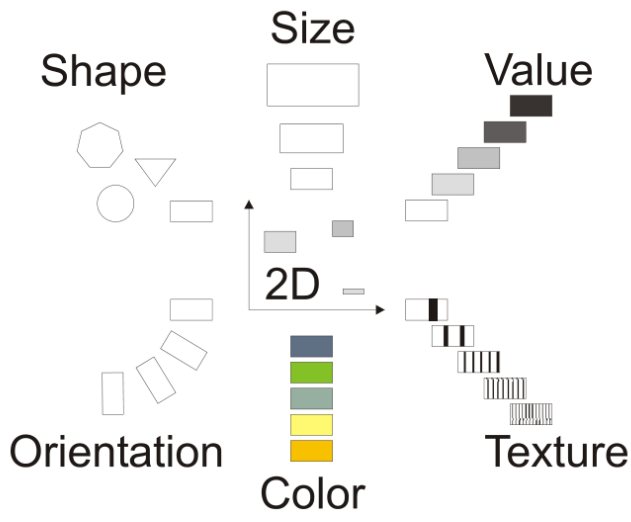
- "An area signifies something on the plane that **has measurable size**.
  This signification applies to the entire area covered by the visible mark."
- an area can change in position but not in size, shape or orientation without making the area itself have a different meaning



From Semiology of Graphics (Bertin)

65

Visual Variables Applicable to Marks

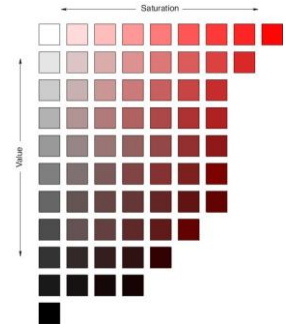


From Semiology of Graphics (Bertin)

Additional Variables for Computers

- **motion**
  - direction, acceleration, speed, frequency, onset, 'personality'

- **saturation**
  - colour as Bertin uses largely refers to hue, saturation != value

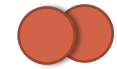


Extending those from Semiology of Graphics (Bertin)

Additional Variables for Computers

- **flicker**
  - frequency, rhythm, appearance

- **depth?** 'quasi' 3D
  - depth, occlusion, aerial perspective, binocular disparity



- **Illumination**

- **transparency**

From Semiology of Graphics (Bertin)

Characteristics of Visual Variables

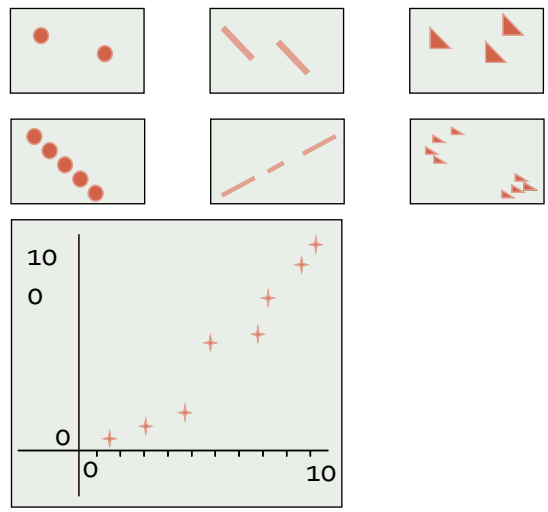
- Selective:
  Can this variable allow us to spontaneously differentiate/isolate items from groups?
- Associative:
  Can this variable allow us to spontaneously group items in a group?
- Ordered:
  Can this variable allow us to spontaneously perceive an order?
- Quantitative:
  Is there a numerical reading obtainable from changes in this variable?
- Length (resolution):
  Across how many changes in this variable are distinctions possible?

From Semiology of Graphics (Bertin)

69

Visual Variable: Position

- ✓ selective
- ✓ associative
- ✓ quantitative
- ✓ order
- ✓ length

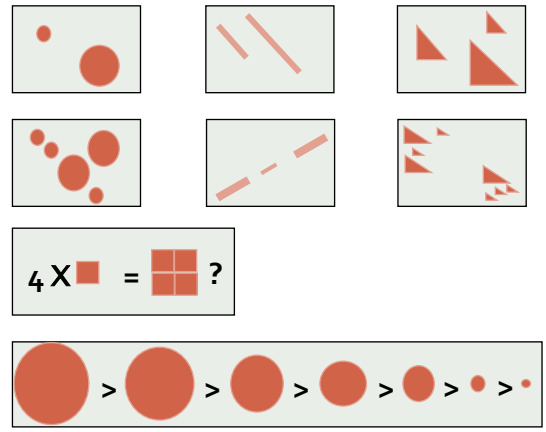


From Semiology of Graphics (Bertin)

70

Visual Variable: Size

- ✓ selective
- ✓ associative
- ≈ quantitative
- ✓ order
- ✓ Length



- theoretically infinite but practically limited
- association and selection ~ 5 and distinction ~ 20

71

Size

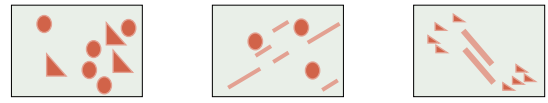
points lines areas

72

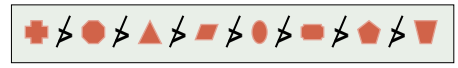
Visual Variable: Shape

≈ • selective

≈ • associative



≠ • ordered



≠ • quantitative

✓ • length
  – infinite

73

Shape

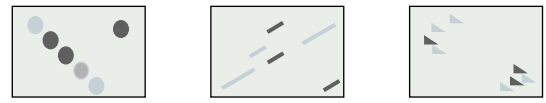
points lines areas

74

Visual Variable: Value

≠ selective

≠ associative



≠ quantitative

✓ order

✓ length

- theoretically infinite but practically limited
- association and selection ~ < 7 and distinction ~ 10

75

Value

points

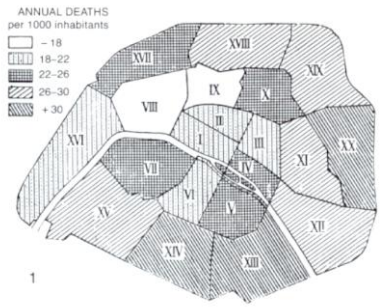
lines

areas

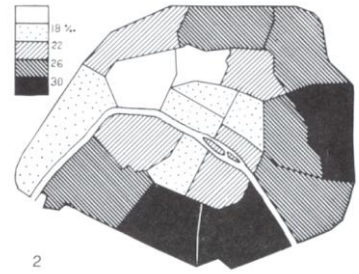
76

Value

- Ordered, cannot be reordered



Values not ordered correctly according to scale
Information has to be read point by point



Values ordered correctly
Image much more useful

annual deaths per 1000 inhabitants, Paris

77

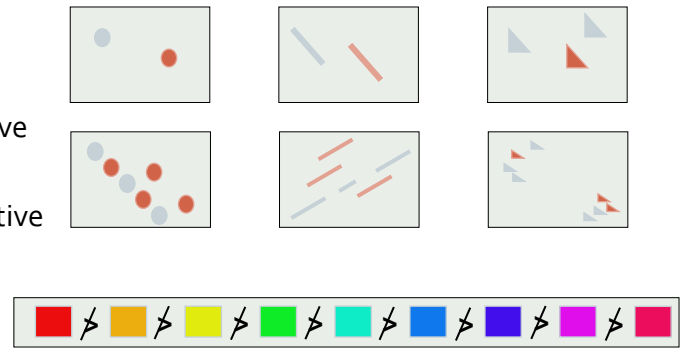
Visual Variable: Colour

✓ selective

✓ associative

≠ quantitative

≠ order



✓ length

- theoretically infinite but practically limited
- association and selection ~ < 7 and distinction ~ 10

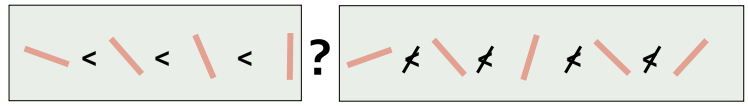
78

Visual Variable: Orientation

- ✓ selective

- ✓ associative

- ≠ quantitative
- ≠ order



- ✓ length
  - ~5 in 2D; ? in 3D

79

Orientation

points lines areas

80

Visual Variable: Texture

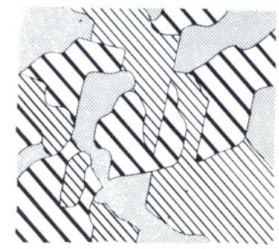
- ✓ selective
- ✓ associative
- ≠ quantitative
- ≠ order
- ✓ length
  - theoretically infinite

81

Texture

points

lines



areas

82

Visual Variable: Motion

- ✓ selective
  - motion is one of our most powerful attention grabbers
- ✓ associative
  - moving in unison groups objects effectively
- ≠ quantitative
  - subjective perception
- ≠ order
- ? length
  - distinguishable types of motion?

83

Motion

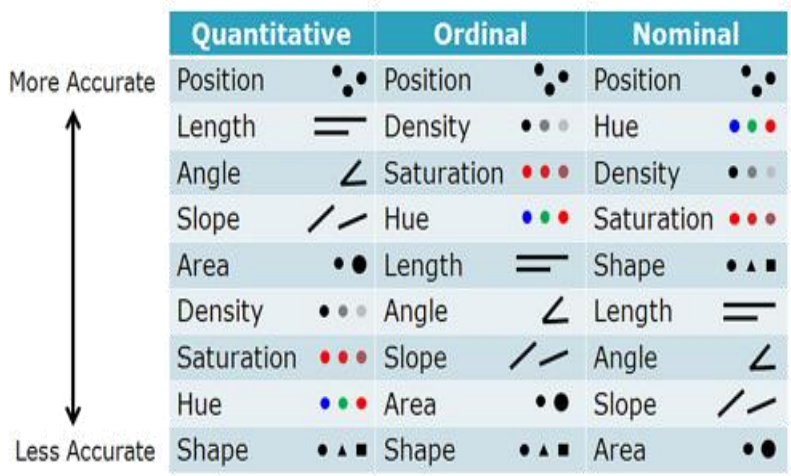
84

Visual Variables

Visual Variable	Selective	Associative	Quantitative	Order	Length
Position	Yes	Yes	Yes	Yes	Dependant on resolution
Size	Yes	Yes	Approximate	Yes	Association: 5; Distinction: 20
Shape	With Effort	With Effort	No	No	Infinite
Value	Yes	Yes	No	Yes	Association: 7; Distinction: 10
Hue	Yes	Yes	No	No	Association: 7; Distinction: 10
Orientation	Yes	Yes	No	No	4
Grain	Yes	Yes	No	No	5
Texture	Yes	Yes	No	No	Infinite
Motion	Yes	Yes	No	Yes	Unknown

Carpendale, 2003 85

Summary



Jacques Bertin refined by Cleveland&McGill then by Card&Mackinlay

Summary

- Now you know the main building blocks are **marks**
- Marks are modified by **visual variables**
- Visual variables have **specific characteristics**
- These characteristics influence how the data will be perceived

Information Visualization

Multidimensional Data

Jean-Daniel Fekete (from John Stasko's slides)
jean-daniel.fekete@inria.fr

Data Sets

- Data comes in many different forms
- Typically, not in the way you want it

- How is stored (in the raw)?
- Heterogeneous data often seen as multiple dimensions of elements extrated by patterns or needs.

Example

- Cars
  - brand
  - model
  - year
  - miles per gallon
  - cost
  - size
  - weights
  - ...

Data Tables

- Often, we take raw data and transform it into a form that is more workable
- Main idea:
  - Individual items are called *cases*
  - Cases have *variables* (attributes)

Data Table Format

		$Case_1$	$Case_2$	$Case_3$	...
Dims	$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
	$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
	$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
	...				

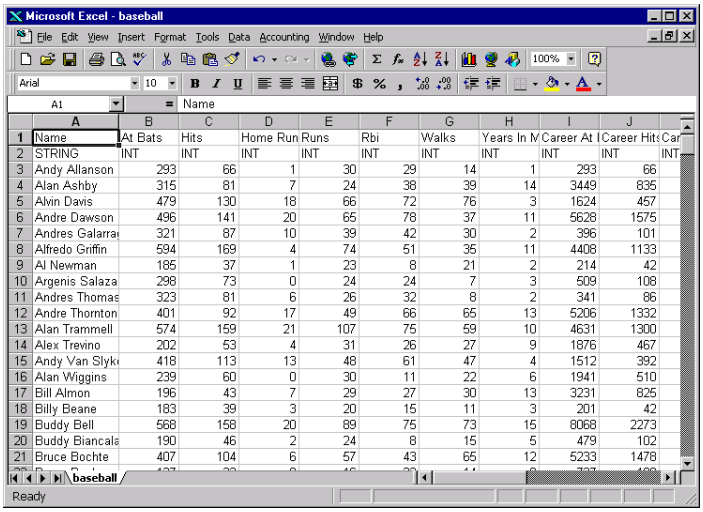
Think of as a function
$f(case_1) = <Val_{11}, Val_{12},...>$

Example

	Mary	Jim	Sally	Mitch	...
SSN	145	294	563	823	
Age	23	17	47	29	
Hair	brown	black	blonde	red	
GPA	2.9	3.7	3.4	2.1	
...					

Example

Baseball statistics



	A	B	C	D	E	F	G	H	I	J
1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In M	Career At I	Career Hits
2	STRING	INT	INT	INT	INT	INT	INT	INT	INT	INT
3	Andy Allanson	293	66	1	30	29	14	1	293	66
4	Alan Ashby	315	81	7	24	38	39	14	3449	835
5	Alvin Davis	479	130	18	66	72	76	3	1624	457
6	Andre Dawson	496	141	20	65	78	37	11	5628	1575
7	Andres Galarra	321	87	10	39	42	30	2	396	101
8	Alfredo Griffin	594	169	4	74	51	35	11	4408	1133
9	Al Newman	185	37	1	23	8	21	2	214	42
10	Argenis Salaza	298	73	0	24	24	7	3	509	108
11	Andres Thomas	323	81	6	26	32	8	2	341	86
12	Andre Thornton	401	92	17	49	66	65	13	5206	1332
13	Alan Trammell	574	159	21	107	75	59	10	4631	1300
14	Alex Trevino	202	53	4	31	26	27	9	1876	467
15	Andy Van Slyk	418	113	13	48	61	47	4	1512	392
16	Alan Wiggins	239	60	0	30	11	22	6	1941	510
17	Bill Almon	196	43	7	29	27	30	13	3231	825
18	Billy Beane	183	39	3	20	15	11	3	201	42
19	Buddy Bell	568	158	20	89	75	73	15	8068	2273
20	Buddy Biancala	190	46	2	24	8	15	5	479	102
21	Bruce Bochte	407	104	6	57	43	65	12	5233	1478

Variable Types

- Three main types of variables
  - N-Nominal (equal or not equal to other values)
    - Example: gender, hair color (blond, brown, black, red)
  - O-Ordinal (obeys < relation, ordered set)
    - Example: soccer leagues, rainbow colors
  - Q-Quantitative (can do math on them)
    - Example: age, photoshop colors
- Sometimes the type depends on the context (see 3 types for color)

Metadata

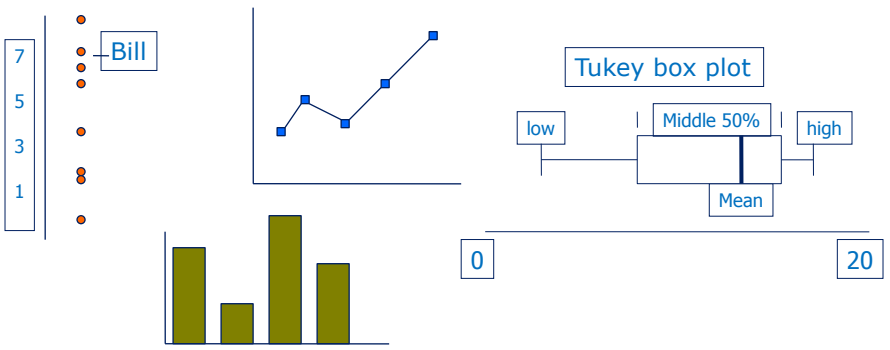
- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example: if variable1 equals "Asia", then variable3 can only be 3, 7 or 16 and variable2 has no more effect on variable3 unless it is zero

How Many Variables?

- Data sets of dimensions 1,2,3 are common
- Number of variables per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data
- Example: www.nationmaster.com

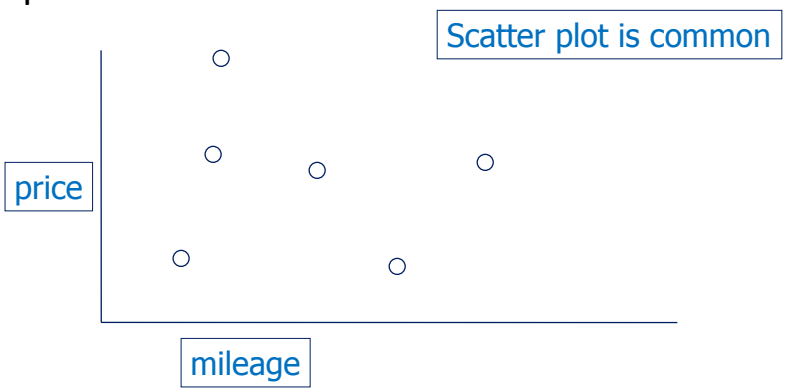
Univariate Data

- Representations



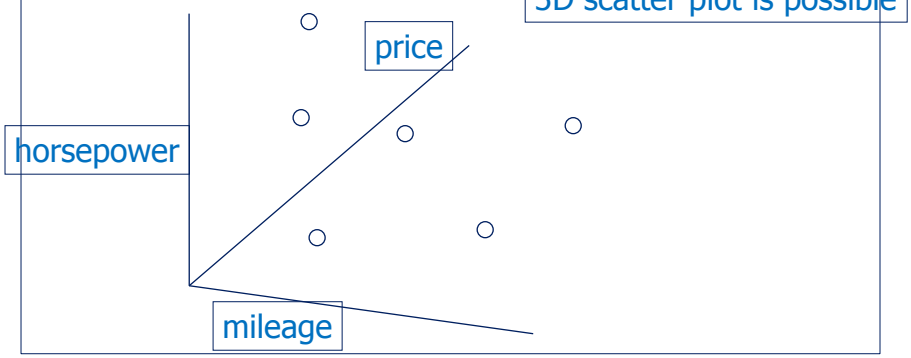
Bivariate Data

- Representations



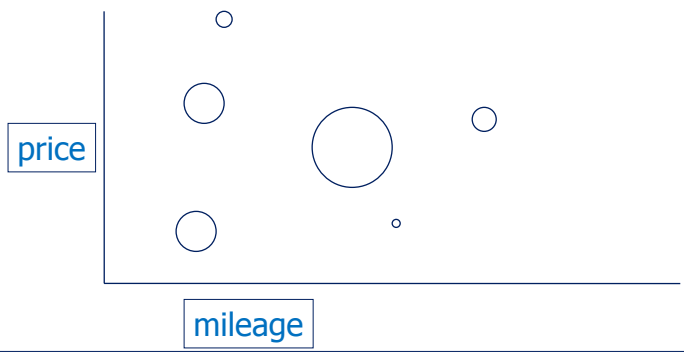
Trivariate Data

- Representation



Alternative View

- Representation based on size instead of z



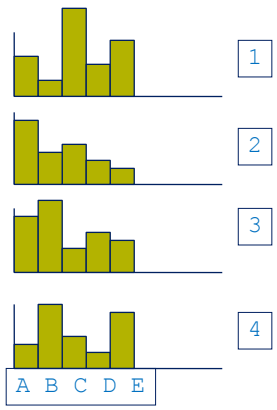
Hypervariate Data

- Number of well-known visualization techniques exist for data sets of 1-3 dimensions
  - line graphs, bar graphs, scatter plots OK
  - We see a 3-D world (4-D with time)
- What about data sets with more than 3 variables?
  - Often the interesting ones

Multiple Views

Give each variable its own display

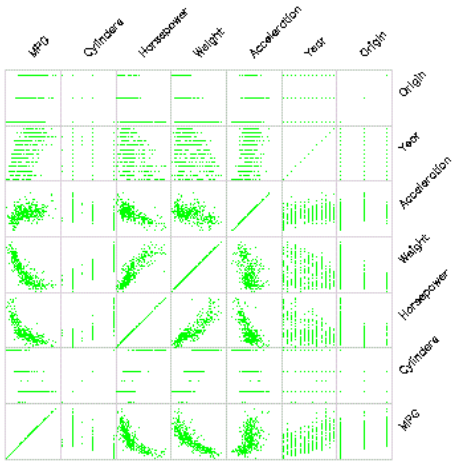
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



Scatterplot Matrix

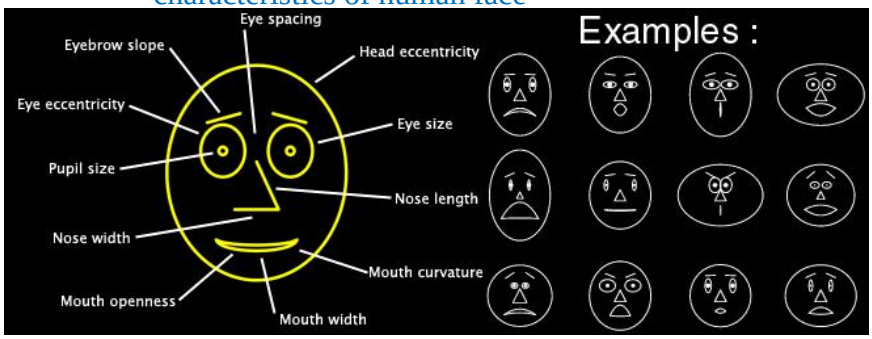
Represent each possible pair of variables in their own 2-D scatterplot

Useful for what?
Misses what?



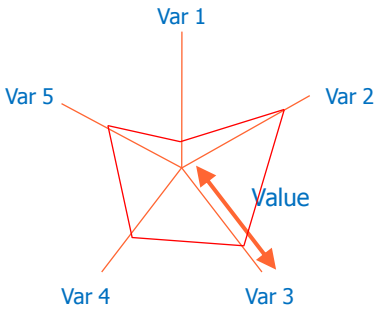
Chernoff Faces

Encode different variables' values in characteristics of human face



Applet: http://www.cs.uchicago.edu/~wiseman/chernoff/
http://hesketh.com/schampeo/projects/Faces/chernoff.html

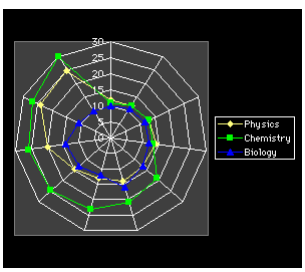
Star Plots

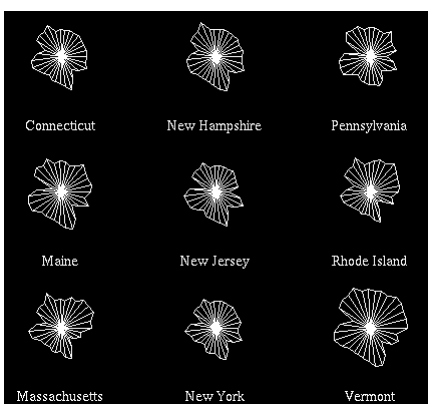


Space out the n variables at equal angles around a circle

Each "spoke" encodes a variable's value

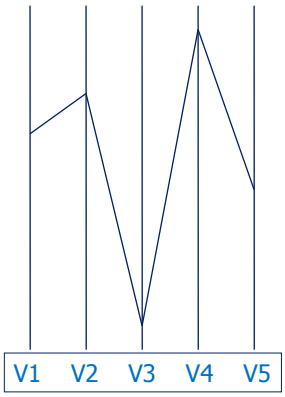
Star Plot examples





http://seamonkey.ed.asu.edu/~behrens/asu/reports/compre/comp1.html

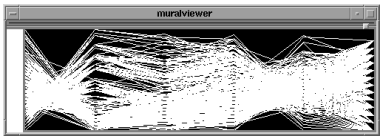
Parallel Coordinates



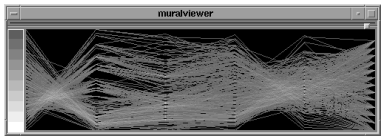
Encode variables along a horizontal row

Vertical line specifies values

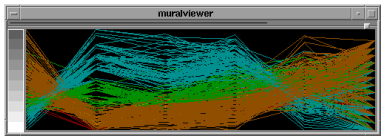
Parallel Coords Example



Basic



Grayscale



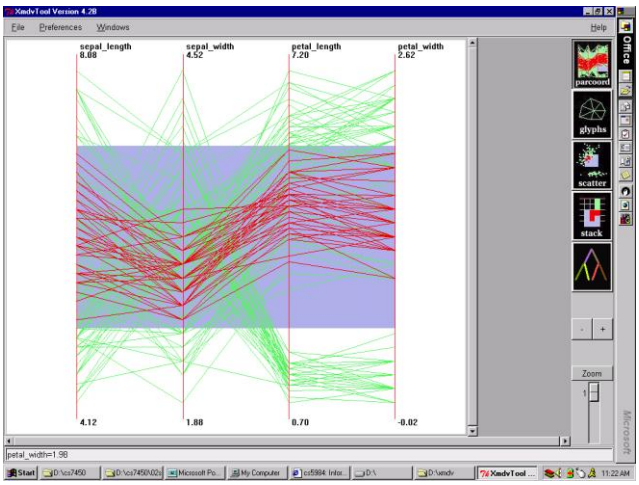
Color

From: Dean F. Jerding and John T. Stasko
http://www.cc.gatech.edu/gvu/softviz/infoviz/information_mural.html

Xmdv

Toolsuite created by Matthew Ward of WPI

Includes parallel coordinate views



Another Technique

- Database of data items, each of n dimensions
- Issue a query that specifies a target value of the dimensions
- Often get back no exact matches
- Want to find near matches

Taken from:
D. Keim, H-P Kriegel, "VisDB Database Exploration Using Multid Vis", *IEEE CG&A*, 1994.

Relevance Factor

- How close an item is to the query
  - Data items have some value that can be numerically quantified
  - Each dimension is some distance away from query item
  - Sum these up for total distance
  - Relevance is inverse of distance

Example

- 5 dimensions, integers 0->255

- Query: 6, 210, 73, 45, 92
- Data item: 8, 200, 73, 50, 91

- Distance: 2 + 10 + 0 + 5 + 1 = 18
- Relevance: 1275 - 18 = 1267

Issues

- What if dimensions are real numbers or text strings?
- What if they're the same type, but of different orders of magnitude?
- Have to define some kind of distance, then a weight function to multiply by

Technique

- Calculate relevance of all data points
- Sort items based on relevance
- Use spiral technique to order the values
- Color items based on relevance

Relevance Colors

High

Low

Empirically established

Spiral Method

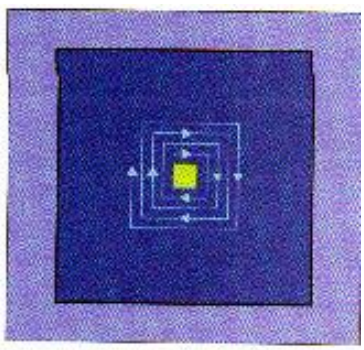


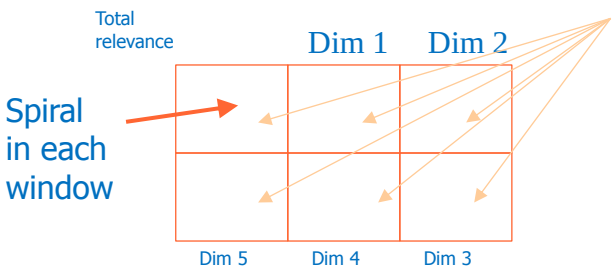
FIGURE 1

Spiral-shaped arrangement of one dimension.

Highest relevance value in center, decreasing values grow outward

Display Methodology

Example: five-dimensional data



Spiral in each window

Same item appears in same place in each window

Items ordered by total relevance

Figure from Paper

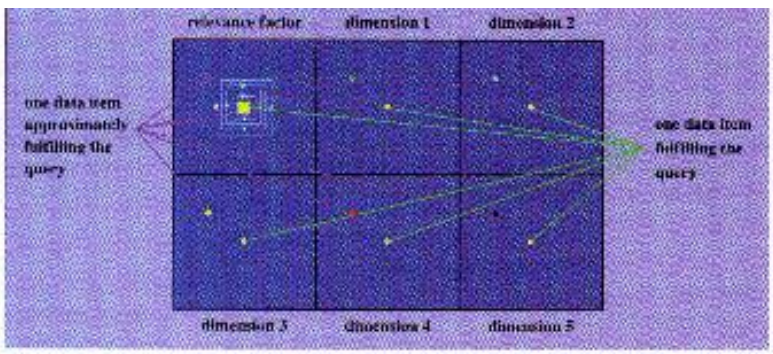
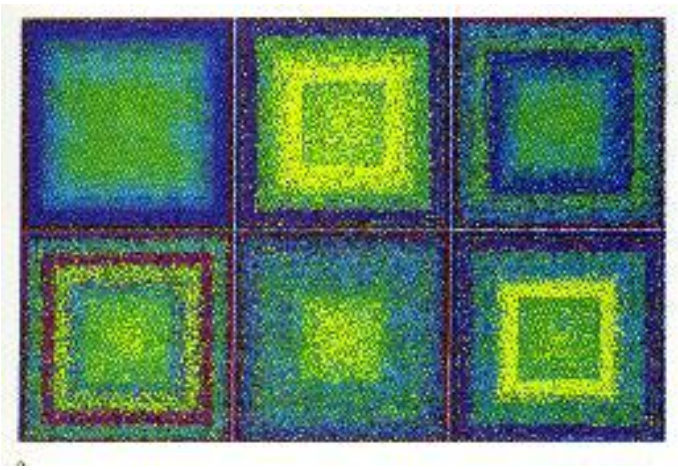


FIGURE 2

Arrangement of windows for displaying five-dimensional data.

Example Display



Alternative

- Grouping arrangement
- Doesn't use multiple windows
- Create all relevance dimensional depictions for an item and group them
- Spiral out the different data items' depictions

Grouping Arrangement

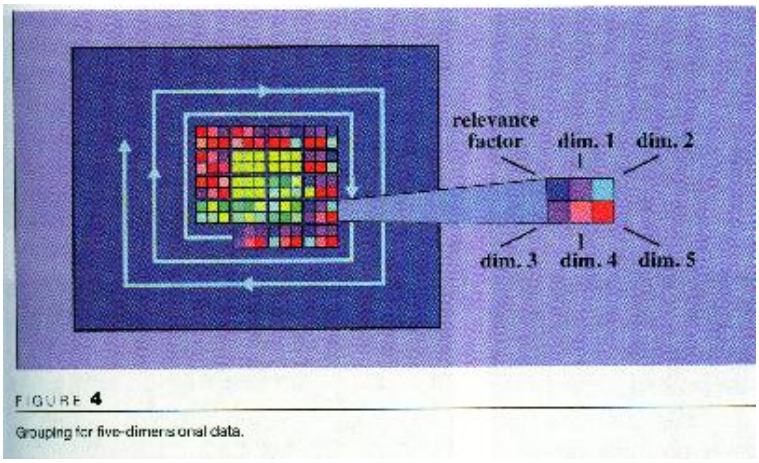
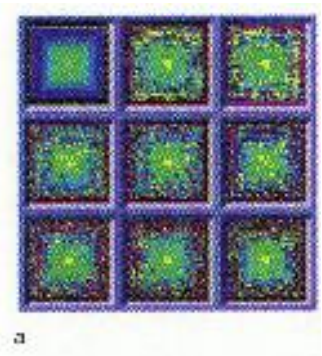


FIGURE 4
Grouping for five-dimensional data.

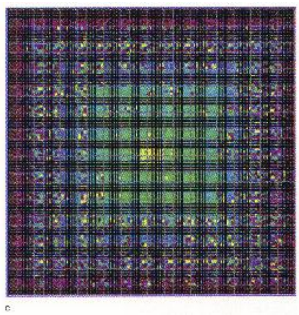
Example Display

8 dimensions

1000 items

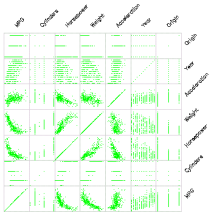


Multi-window

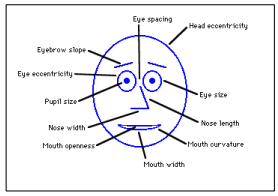


Grouping

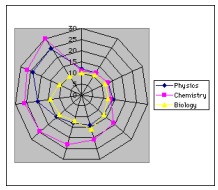
Overview



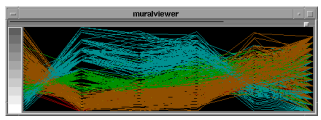
Scatterplot Matrix



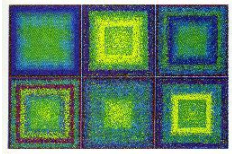
Chernoff Faces



Star Plots



Parallel Coordinates

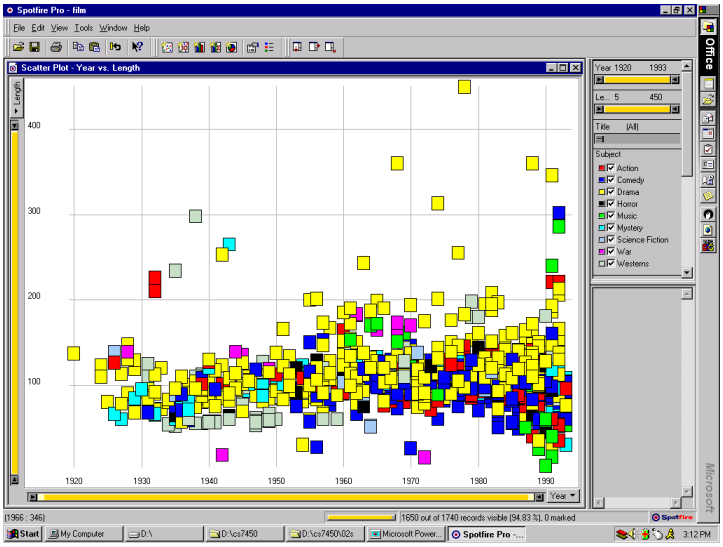


Spiral plots

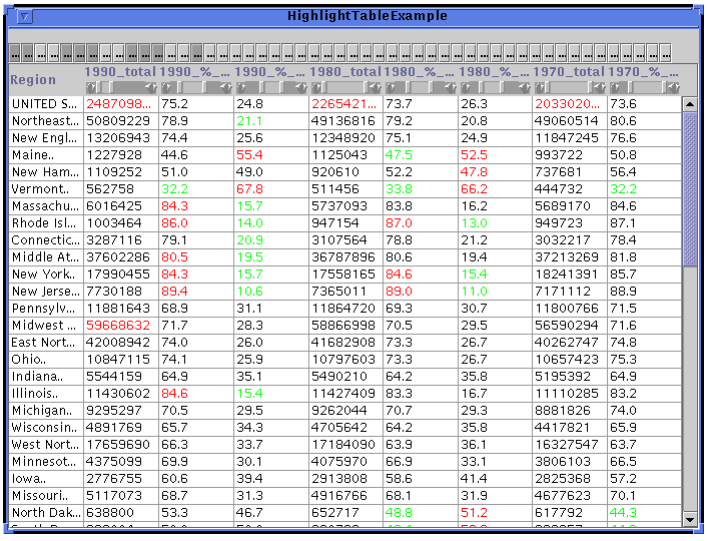
MultiVariate Visu Tools

INTERACTION is the key !

SpotFire



Highlighted Dynamic Table Viewer (HiDTV)



HighlightTableExample

Region	1990_total	1990_%_...	1990_%_...	1980_total	1980_%_...	1980_%_...	1970_total	1970_%_...
UNITED S...	2487098...	75.2	24.8	2265421...	73.7	26.3	2033020...	73.6
Northeast...	50809229	78.9	21.1	49136816	79.2	20.8	49060514	80.6
New Engl...	13206943	74.4	25.6	12348920	75.1	24.9	11847245	76.6
Maine..	1227928	44.6	55.4	1125043	47.5	52.5	993722	50.8
New Ham...	1109252	51.0	49.0	920610	52.2	47.8	737681	56.4
Vermont..	562758	32.2	67.8	511456	33.8	66.2	444732	32.2
Massachu...	6016425	84.3	15.7	5737093	83.8	16.2	5689170	84.6
Rhode Isl...	1003464	86.0	14.0	947154	87.0	13.0	949723	87.1
Connectic...	3287116	79.1	20.9	3107564	78.8	21.2	3032217	78.4
Middle At...	37602286	80.5	19.5	36787896	80.6	19.4	37213269	81.8
New York..	17990455	84.3	15.7	17558165	84.6	15.4	18241391	85.7
New Jerse...	7730188	89.4	10.6	7365011	89.0	11.0	7171112	88.9
Pennsylv...	11881643	68.9	31.1	11864720	69.3	30.7	11800766	71.5
Midwest ...	59668632	71.7	28.3	58866998	70.5	29.5	56590294	71.6
East Nort...	42008942	74.0	26.0	41682908	73.3	26.7	40262747	74.8
Ohio..	10847115	74.1	25.9	10797603	73.3	26.7	10657423	75.3
Indiana..	5544159	64.9	35.1	5490210	64.2	35.8	5195392	64.9
Illinois..	11430602	84.6	15.4	11427409	83.3	16.7	11110285	83.2
Michigan..	9295297	70.5	29.5	9262044	70.7	29.3	8881826	74.0
Wisconsin..	4891769	65.7	34.3	4705642	64.2	35.8	4417821	65.9
West Nort...	17659690	66.3	33.7	17184090	63.9	36.1	16327547	63.7
Minnesot...	4375099	69.9	30.1	4075970	66.9	33.1	3806103	66.5
Iowa..	2776755	60.6	39.4	2913808	58.6	41.4	2825368	57.2
Missouri..	5117073	68.7	31.3	4916766	68.1	31.9	4677623	70.1
North Dak...	638800	53.3	46.7	652717	48.8	51.2	617792	44.3

InfoZoom 1

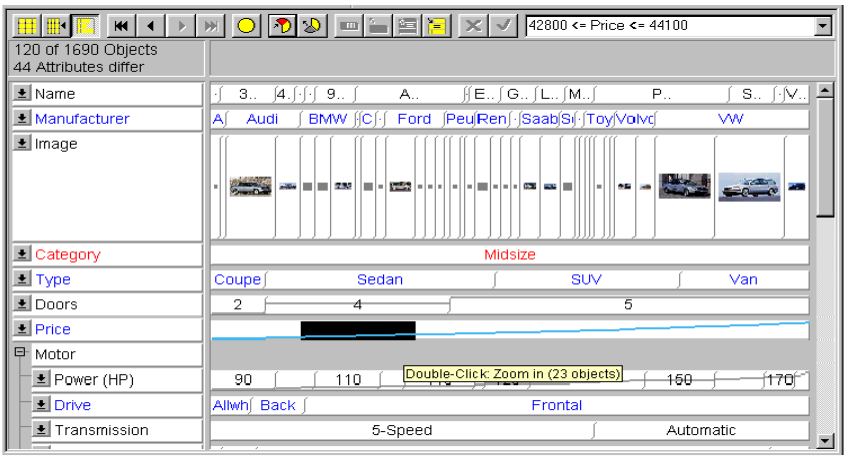
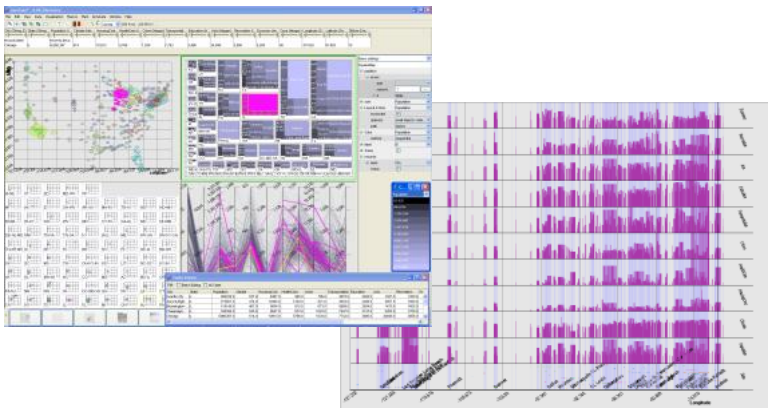


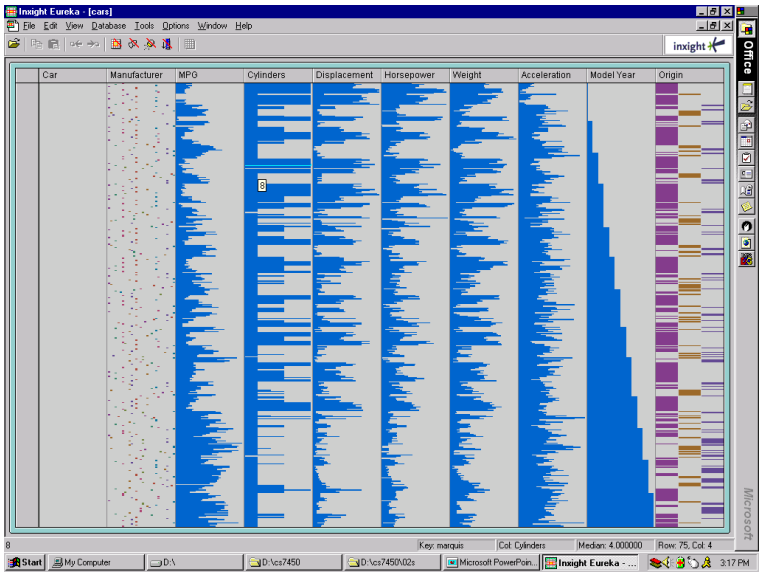
Figure 1: InfoZoom – Overview Mode

ILOG Discoveries

- http://www2.ilog.com/preview/Discovery/



Eureka / TableLens



Conclusion

- Few effective visual representations for multidimensional data
- Some more require a lot of training
  - Not appropriate for beginners or casual users
- Better interaction facilitates learning and navigation

Information Visualization: Trees and Hierarchies

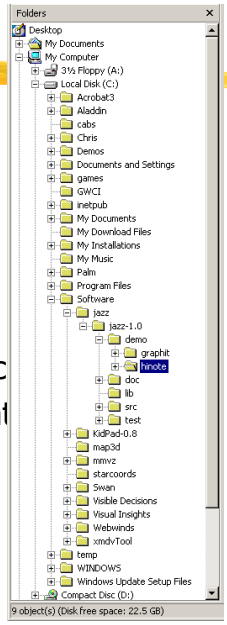
Jean-Daniel Fekete
INRIA Saclay – Île-de-France
Projet AVIZ

Trees (Hierarchies)

- What is a tree?
  - DAG, one parent per node
  - Items + structure (nodes + links)
  - Table model: Add parent pointer attribute
- Examples
  - filesystem, family, classification/taxonomy, org charts, toc, data structures, menus
- Tasks
  - All previous tasks plus structure-based tasks:
  - Find descendants, ancestors, siblings, cousins
  - Overall structure, height, breadth, dense/sparse areas

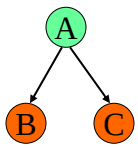
Tree Visualization

- Example: Outliner
- Why is tree visualization hard?
  - Structure AND items
  - Structure harder, consumes more spac
  - Data size grows very quickly (exponent
    - #nodes = b^{height}

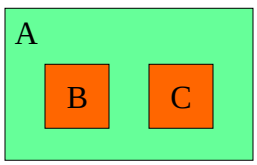


2 Approaches

- Connection (node & link)
  - outliner



- Containment (node in node)
  - Venn diagram

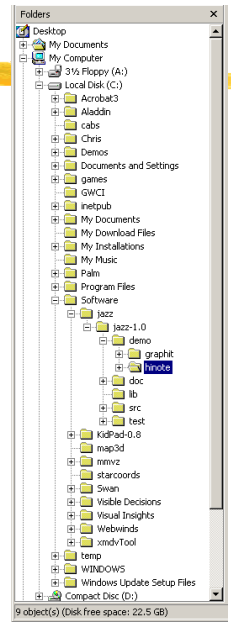


Tree Properties

- Structure vs. attributes
  - Attributes only (multi-dimensional viz)
  - Structure only (1 attribute, e.g. name)
  - Structure + attributes
- Branching factor
- Fixed level, categorical

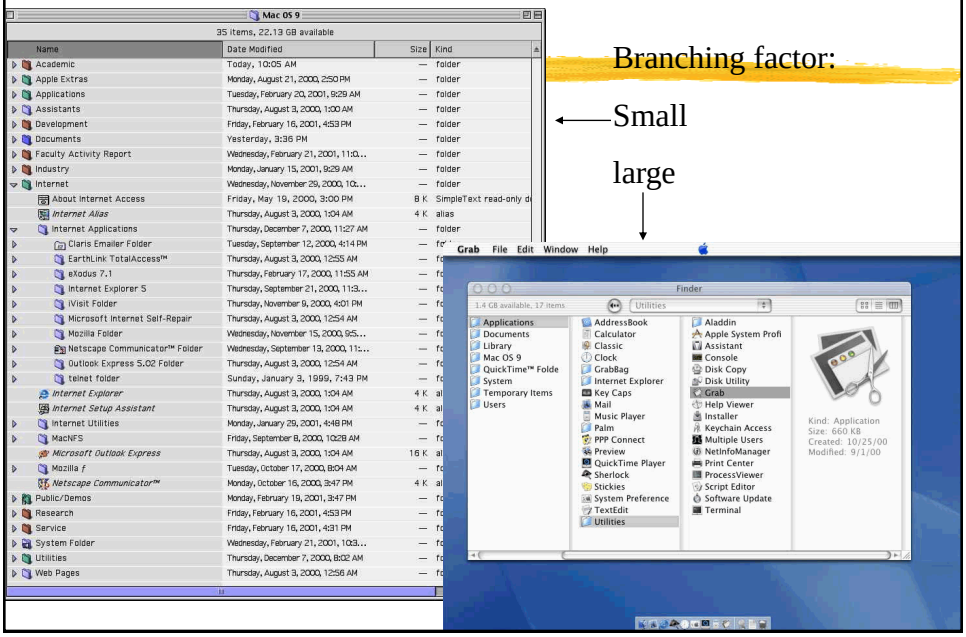
Outliner

- Good for directed search tasks
- Not good for learning structure
- No attributes
- Apx 50 items visible
- Lose path to root for deep nodes



- cant see all the tree structure
- Scroll bar suck
- Structure only
- Lost screen space
- 50 nodes

Mac Finder



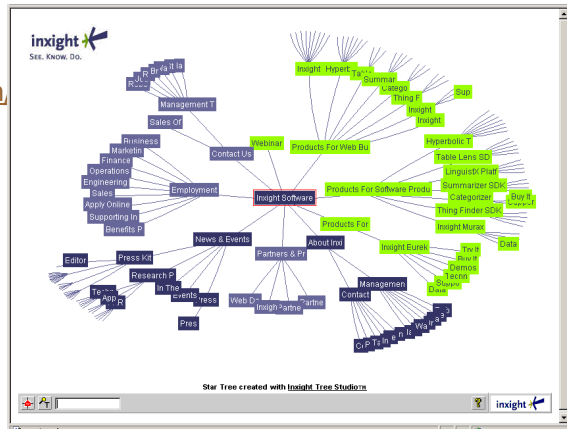
Branching factor:

Small

large

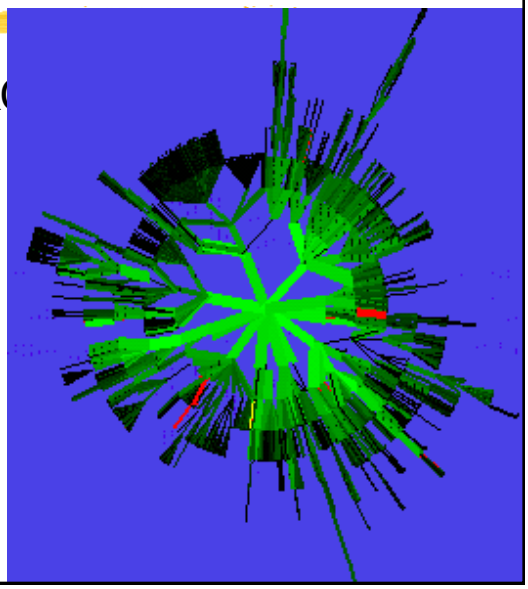
Hyperbolic Trees

- Rao, "Hyperbolic Tree"
  - David, Harsha
- http://startree.inxight.com/
- Xerox PARC
- Inxight



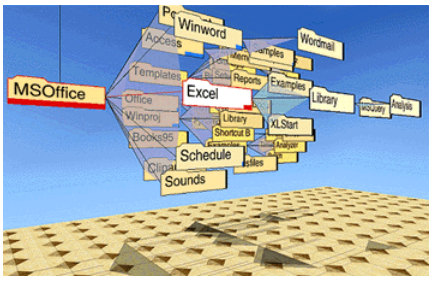
Disk Tree

- Ed Chi, Xerox PAR



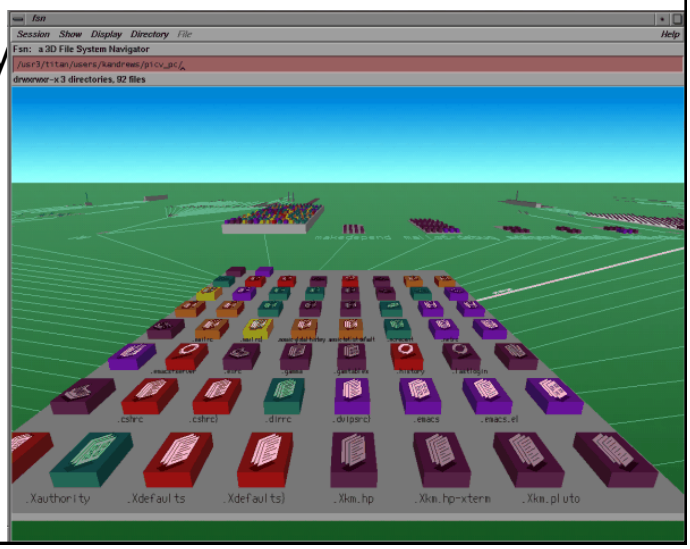
Cone Trees

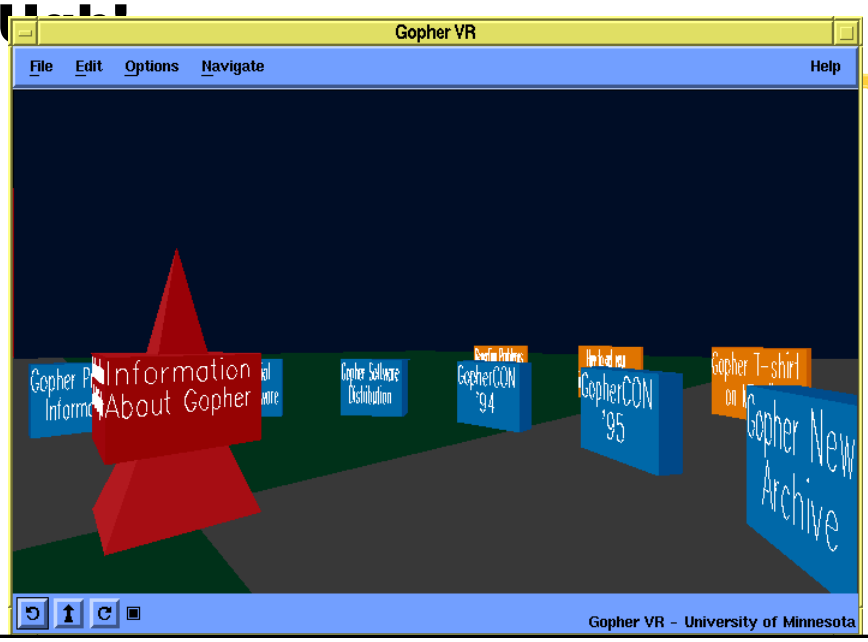
- Robertson, "ConeTrees"
  - Anuj, Atul
- Xerox PARC



FSN

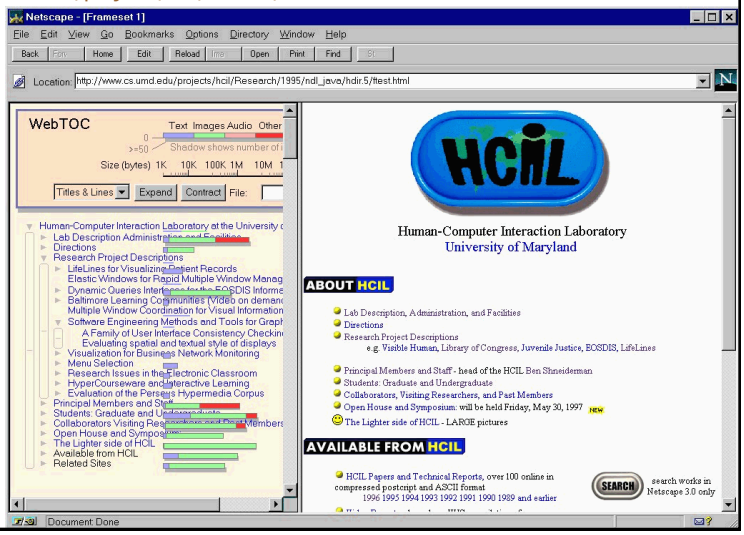
- SGI file sy
- Jurassic Park





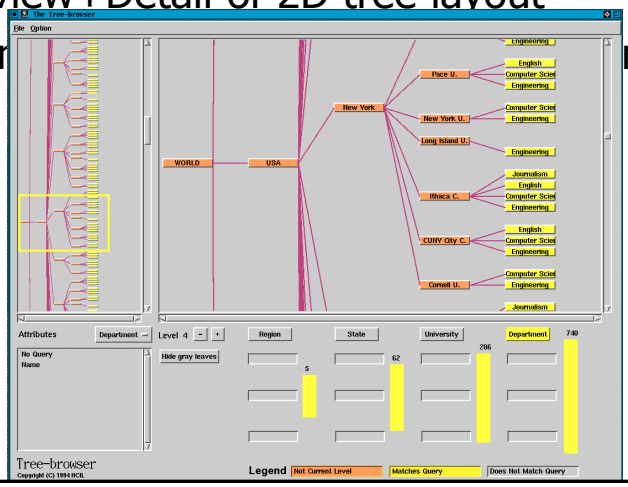
WebTOC

- Website map: Outliner + size attributes
- http://www.cs.umd.edu/projects/hcil/webtoc/fhcil.html

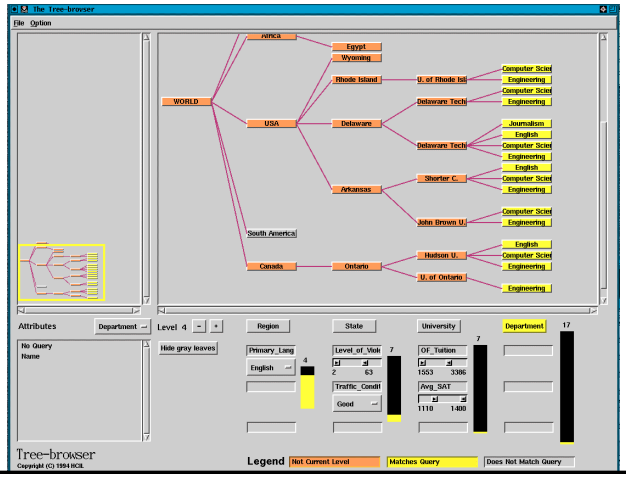


PDQ Trees

- Overview+Detail of 2D tree layout
- Dynar ning

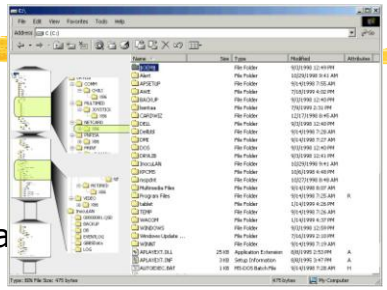


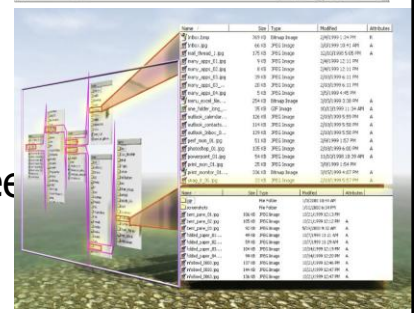
PDQ Trees



Hard Problems

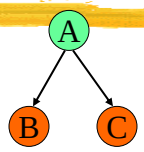
- Multiple foci
  - Robertson, Microsoft Resea
- Polyarchies:
  multiple inter-twined tree



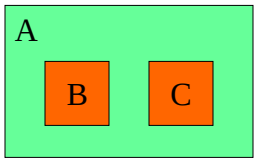


2 Approaches

- Connection (node & link)
  - Outliner, …

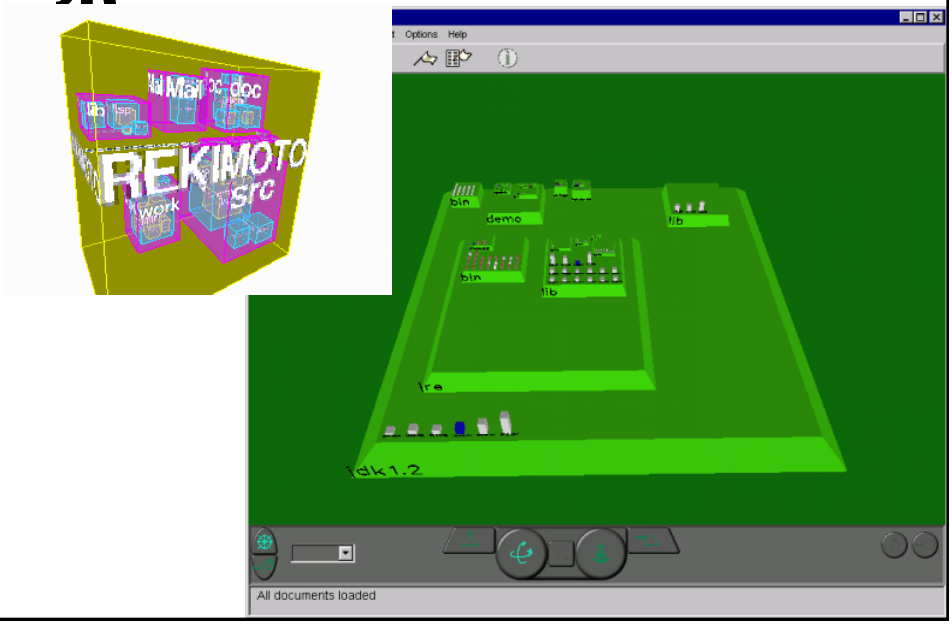


- Containment (node in node)
  - Venn diagram



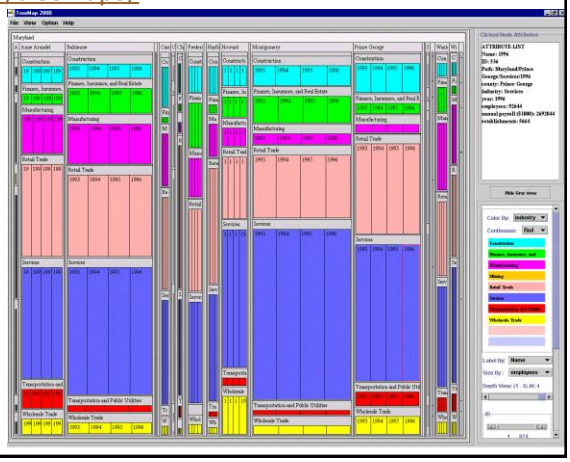
- Structure vs. attributes
  - Attributes only (multi-dimensional viz)
  - Structure only (1 attribute, e.g. name)
  - Structure + attributes

3D



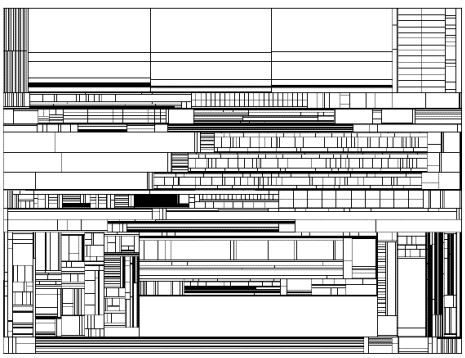
Treemaps

- Shneiderman, "Treemaps"
  - Vishal, Jeevak
- http://www.cs.umd.edu/hcil/treemap3/
- Maryland

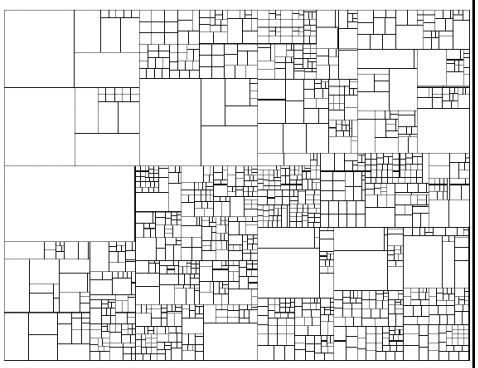


Squared Treemaps

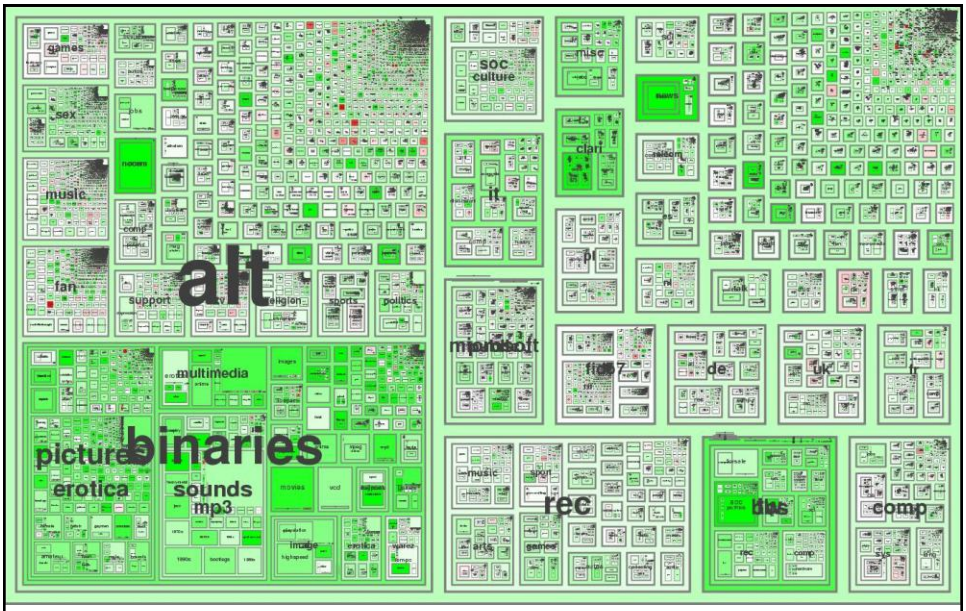
- Wattenberg
- Van Wijk



(a) File system



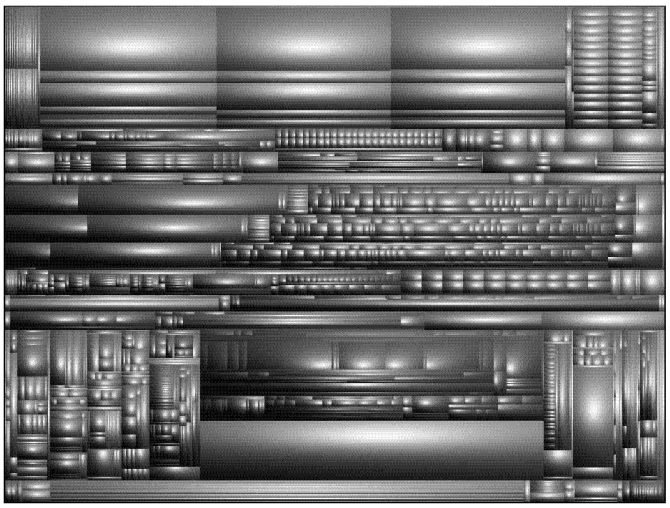
(a) File system



- http://www.research.microsoft.com/~masmith/all_map.jpg

Cushion Treemaps

- Van Wijk
- http://www.win.tue.nl/sequoiaview/

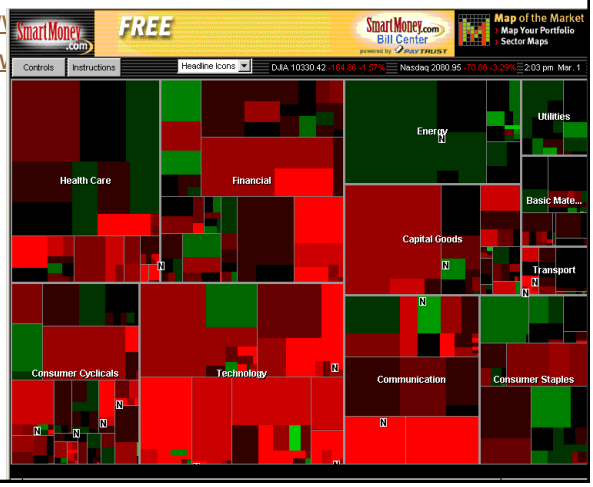


Dynamic Queries

- http://www.cs

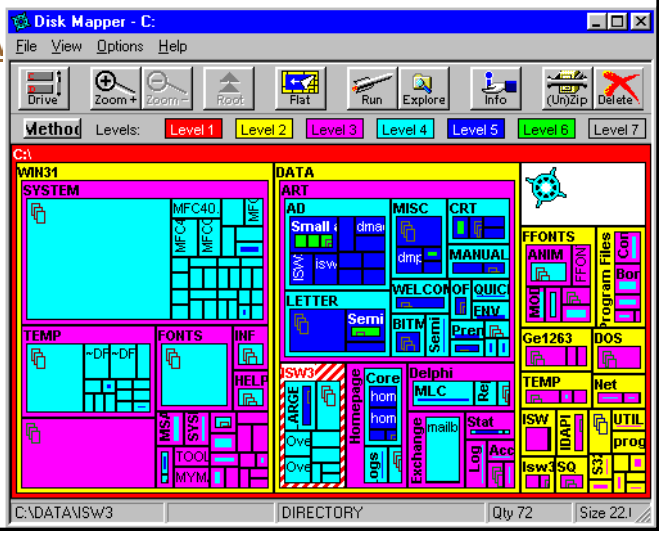
Treemaps on the Web

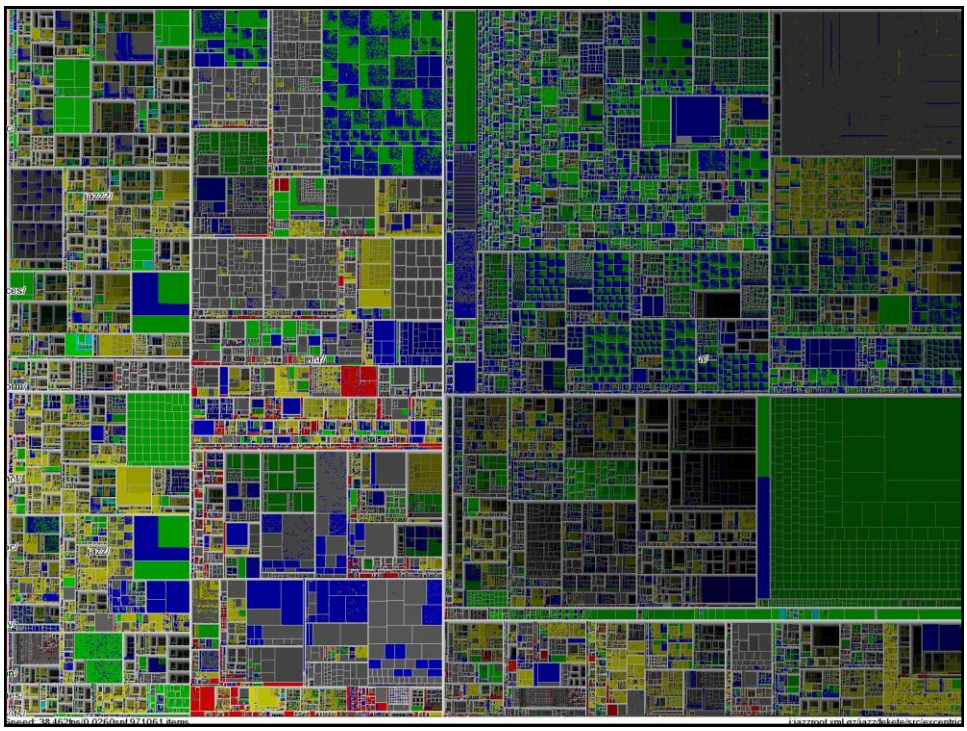
- Map of the Market: http://www.smartmoney.com/marketmap/
- People Map: http://ww
- Coffee Map: http://ww



DiskMapper

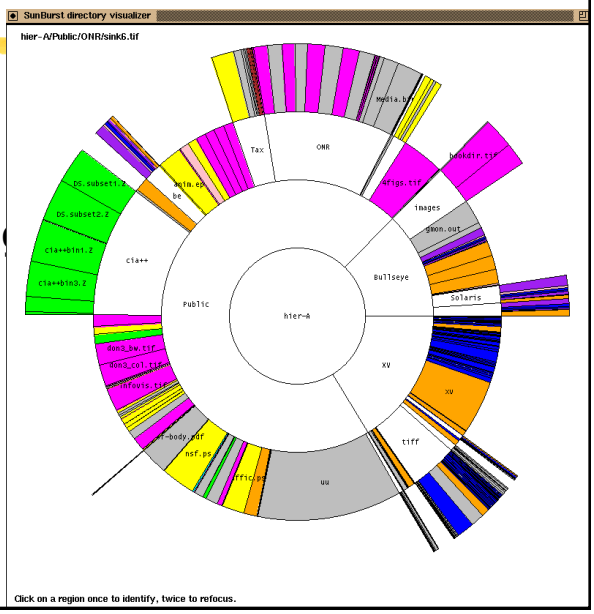
- http://www





Sunburst

- Stasko, GaTech
- Radial layout
- Animated zoomin



Sunburst vs. Treemap

- + Faster learning time: like pie chart
- + Details outward, instead of inward
- + Focus+context zooming

- - Not space filling
- - More space used by non-leaves
  - All leaves on 1-D space, perimeter
  - Treemap: 2-D space for leaves
- - Smaller scale?

Conclusion

- Two representations plus variants
- Don't invent your own!
- Scale to about 1 million items for Treemaps, much smaller for Node-Link diagrams
  - But possibilities for multi-scale interactions

Network Visualization

Jean-Daniel Fekete
INRIA Saclay – Île-de-France

Formal definition

- A graph is:
  - A set of Vertices V={v_i}
  - A set of edges E={e_j} with e=(v_s,v_d) ∈ VxV
  - When the order of the couple in E is meaningful, the graph is **directed**, otherwise, it is **undirected**
  - A graph is a mapping of V into V
  - From there, we can define several measures or intrinsic properties on a graph

2

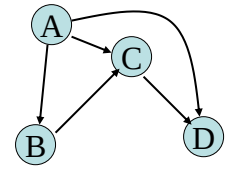
Properties of a Graph

- From a graph structure, we can compute the following properties:
  - Degree (input/output) of a vertex
  - Distance between two vertices (min number of hops from one to the other)
  - Connected components (weak/strong)
  - Diameter
  - Centrality of a vertex
  - Density of a graph

3

Two classical visual representations of a Graph

- Node Link Diagram
  - A node represents a vertex
  - A link represents an edge
- Adjacency Matrix
  - A line is a (source) vertex
  - A column is a (destination) vertex
  - A the intersection, the cell represents the edge



↗	A	B	C	D
A		X	X	X
B			X	
C				X
D				

4

Examples of Tasks on Graphs

- The main tasks related to graphs are:
  - **Vertices**: find their degree, if they are isolated (degree=0), source only (no incoming edge), sink (no outgoing edges), neigbors
  - **Paths**: shortest paths, all paths, cycles
  - **Sub-Graph:** connected components, etc.
- The list is unbounded, all the application domains add more useful tasks. However, there is a bounded list of very generic tasks.

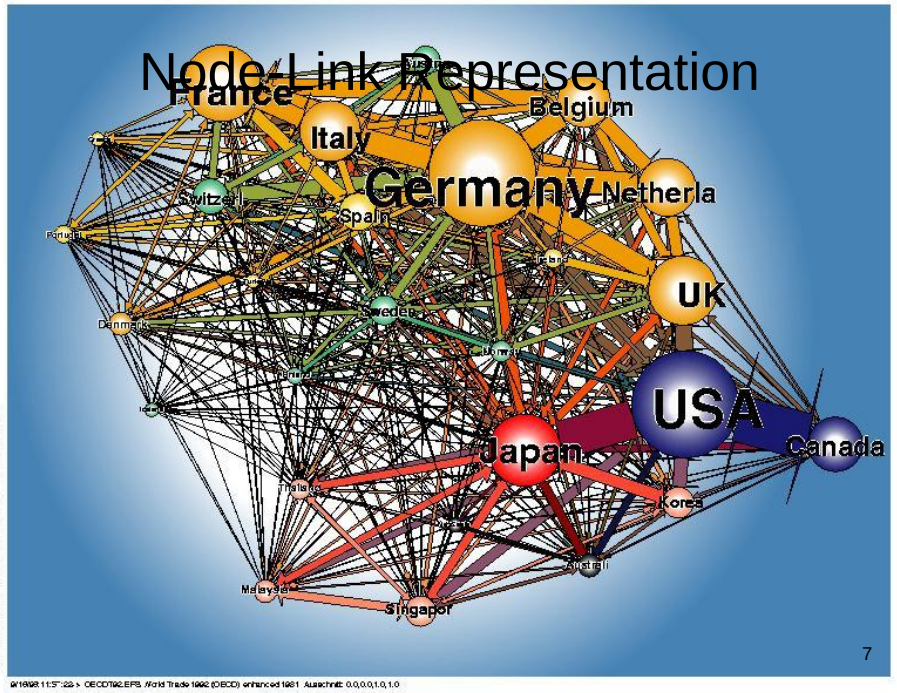
5

Attributes of a Graph

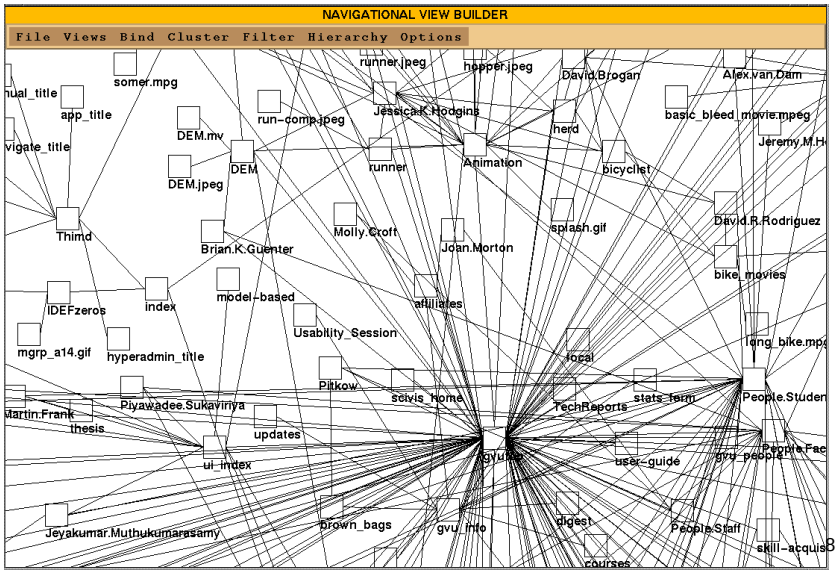
- The Graph structure only defines a **topology**
- Attributes can also be associated to vertices and edges (I call the attributed graph a **Network)**
  - Examples: a label, weight, length, count…
- With attributes, new measures can be computed on a Network:
  - Shortest path according to edge length, cycles with constraints, etc.
- New tasks related to attributes are added
  - Shortest cycle going through the cities with the largest population

6

Node-Link Representation



What is the problem?



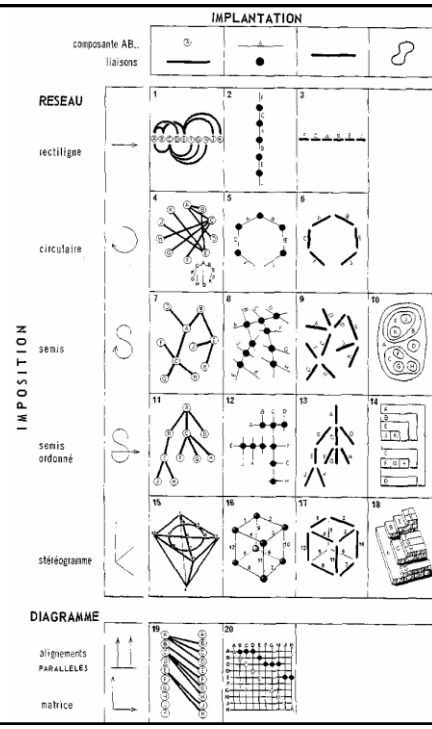
Issues

- Graph Layout
- Scalability
- Navigation

Beware of the hairball!

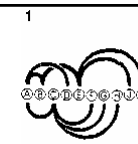
9

Bertin 67

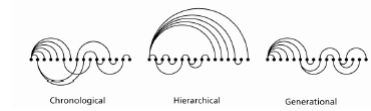


10

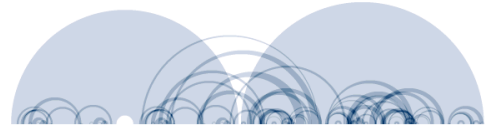
Linear layout



- **THREAD ARCS: An Email Thread Visualization** (Bernard Kerr, IBM)



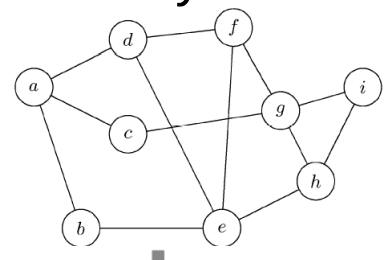
- Arc Diagrams: The Shape of Song (Martin Wattenberg, IBM)
- http://www.bewitched.com/



11

Linear Layout Theory

- Order vertices so that a structure appears
  - Layout φ
- Some well-known objective functions (Diaz et al. 02):
  - Bandwidth
  - MinLA
  - Cut Width
  - Min Cut
  - Sum Cut



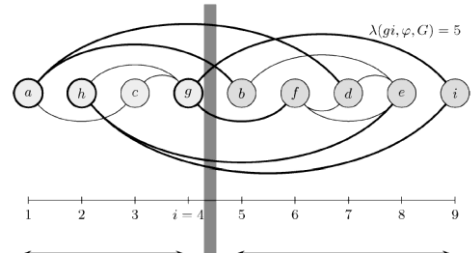
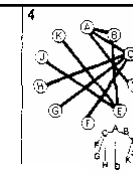


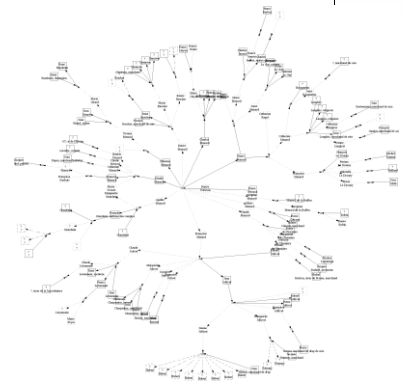
Fig. 1. A graph G together with some layout measures and a graphical representation of the layout $\varphi = \{(a, 1), (b, 5), (c, 3), (d, 7), (e, 8), (f, 6), (g, 4), (j, 9), (h, 2)\}$.

12

Circular Layout



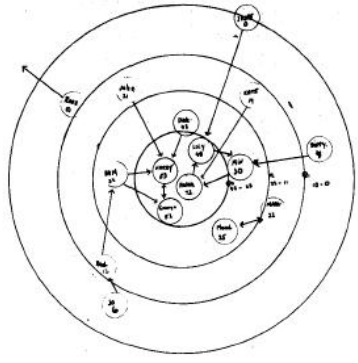
- Egocentric Social Networks
- Can be ordered too

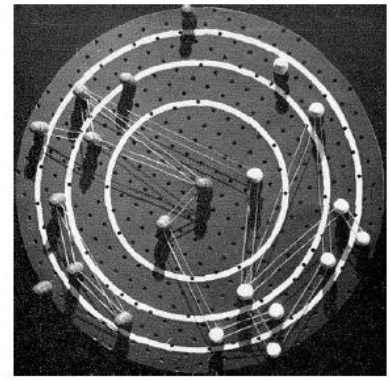


13

centrality visualization

Northway (1940)

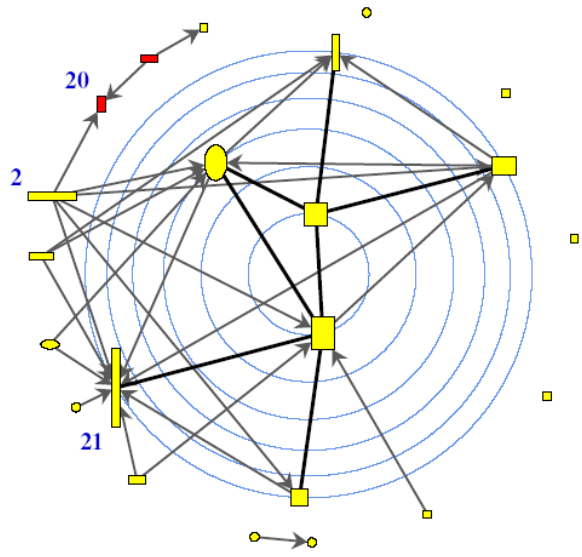




14

centrality visualization

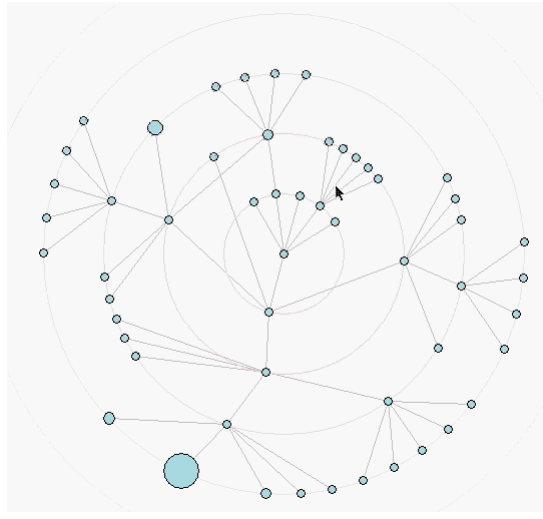
network of organizations involved in policy making



15

Animation of a circular network

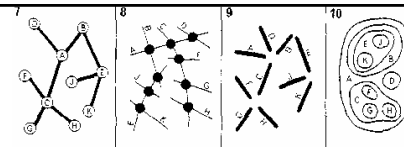
[Animated Exploration of Graphs with Radial Layout, InfoVis'01]



16

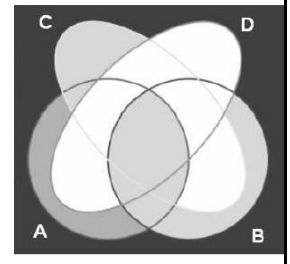
http://www.youtube.com/watch?v=OPX5iGro_lA

Free Layout



- Order is not meaningful, only proximity is
- Force-directed methods
- Treemaps and Venn Diagrams

for $t \leftarrow 1$ **to** *ITERATIONS* **do**
  for $v \in V$ **do**
    $D \leftarrow \sum_{u:\{u,v\}\notin E} f_{\text{rep}}(p_u, p_v) + \sum_{u:\{u,v\}\in E} f_{\text{spring}}(p_u, p_v)$;
    $p_v \leftarrow p_v + \delta \cdot D$



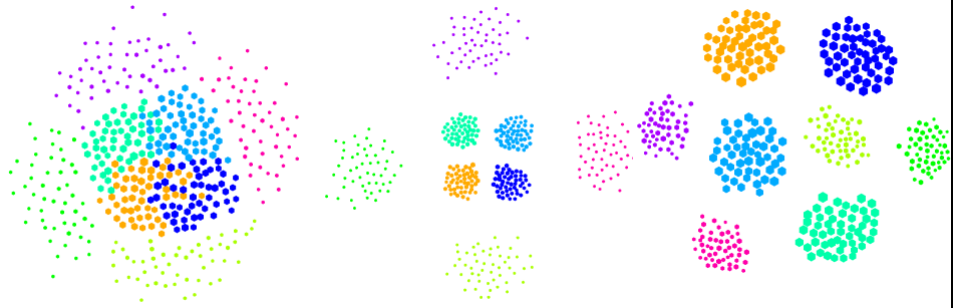
17

Force-directed Layout

- Spring Model: f_{rep} linear, f_{att} linear
  - Good space filling when the network is sparse and homogeneous
- Kamada&Kawai : f_{rep} linar, f_{att} function of the Graph-theoretical distrance
  - Shows the shortest paths as quasi straight lines
- Fruchterman&Reingold : f_{att} quadratic
  - Fills the space better for varying degrees
- Davidson&Harel : f_{rep} linear, f_{att} linear*degree
  - Better space filling for heterogenous graphs
- Hall/Spectral
  - Very fast but works best with grids
- LinLog : f_{rep} linear, f_{att} linear+log
  - Show clusters

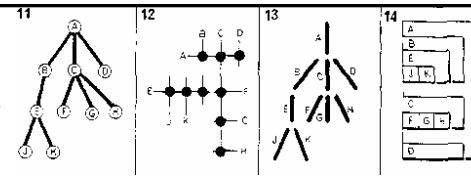
18

Example

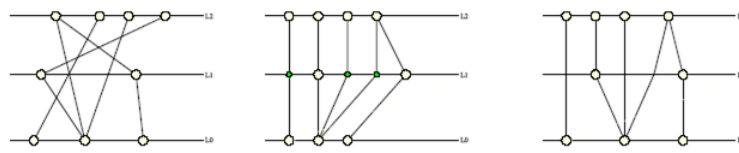


Fruchterman-Reingold model | Node-repulsion LinLog | Edge-repulsion LinLog mode

Ordered layouts



- One dimension is orderable (e.g. genealogical trees)
  - The other can be partially ordrer
- Sugiyama algorithm
- Improved by C. North [A Technique for Drawing Directed Graphs.]

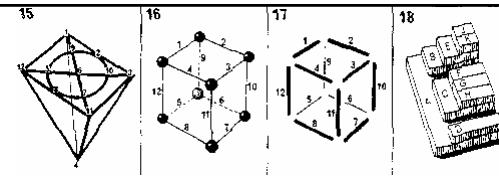


vertical ordering
layer assignment

horizontal ordering
crossing reduction

coordinate assignment
spacing and straightness

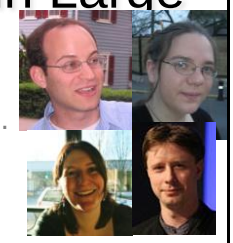
3D Networks



- 3D methods usually extend to 3D naturally
- Navigation is then required

21

Topology-Aware Navigation in Large Networks [CHI 09]

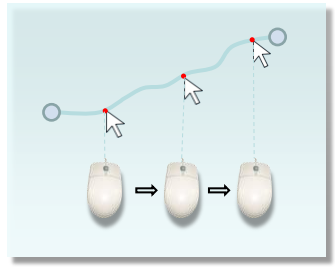


T. Moscovich, F. Chevalier, N. Henry, E. Pietriga, J.-D.

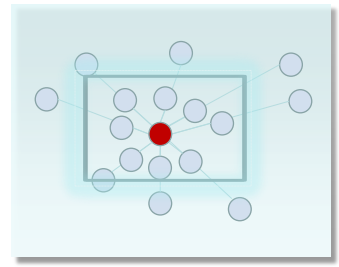
Aviz Visual Analytics Project in|situ| INRIA MICROSOFT RESEARCH INRIA JOINT CENTRE

Further exploiting topological information

Link sliding



Bring'n Go

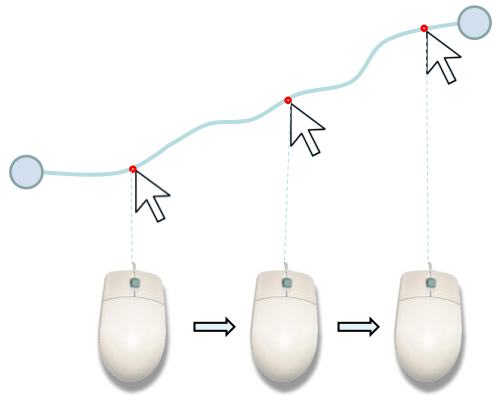


Navigating in large networks

Link Sliding

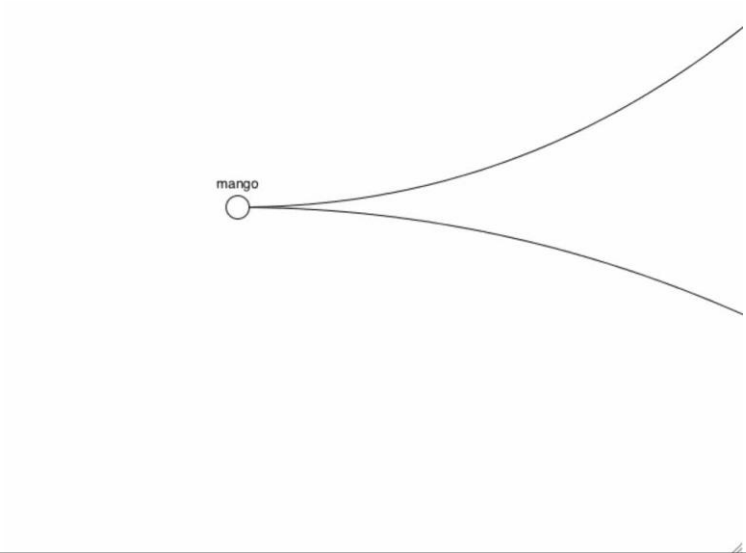
- Motion constrained on the links

  linear motion -> curvilinear motion



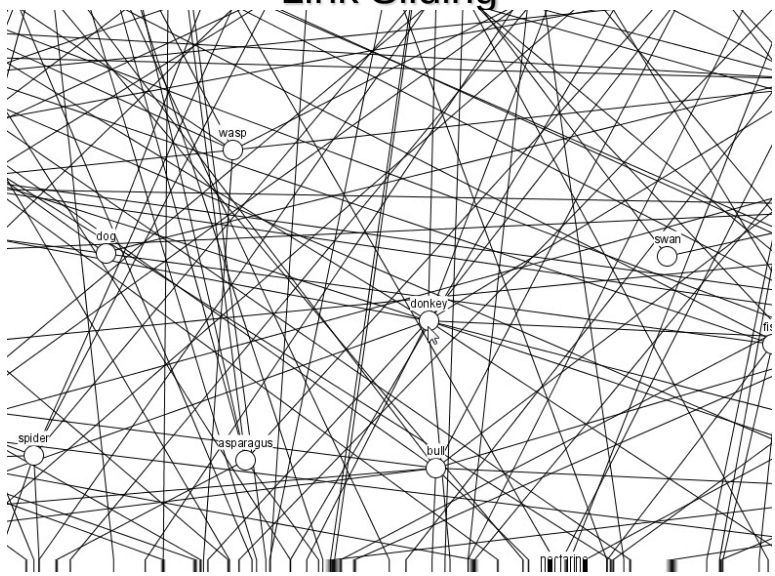
Navigating in large networks

Link Sliding



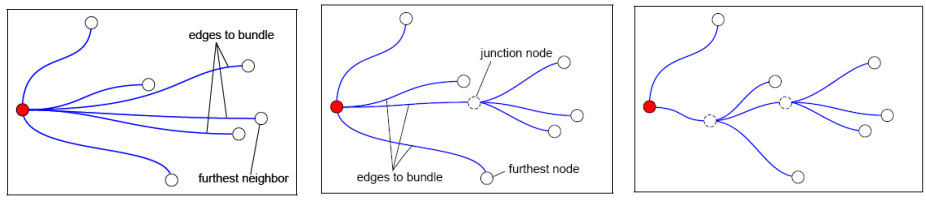
Navigating in large networks

Link Sliding

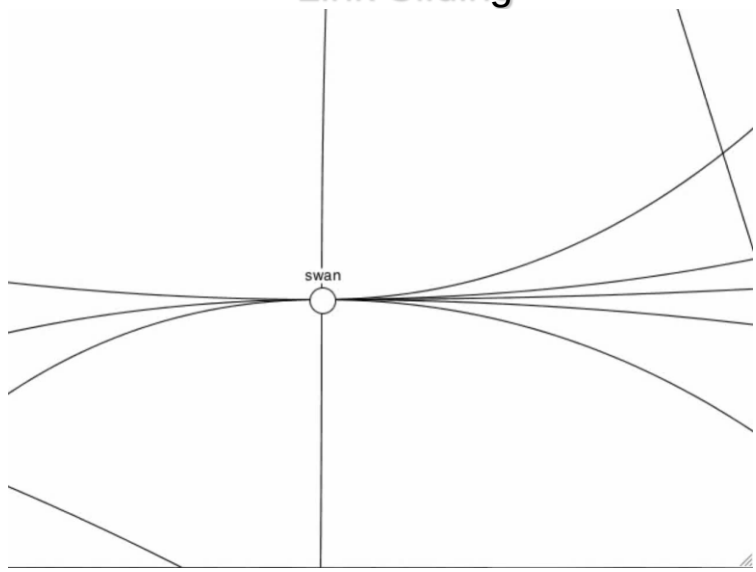


Navigating in large networks
Link Sliding

- **Edge bundling**
  Facilitate link selection

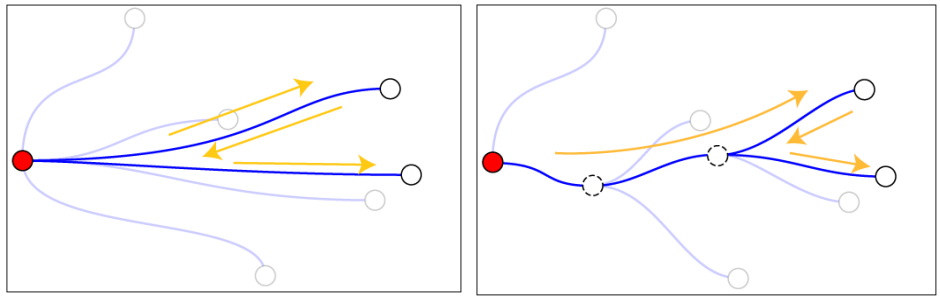


Navigating in large networks
Link Sliding



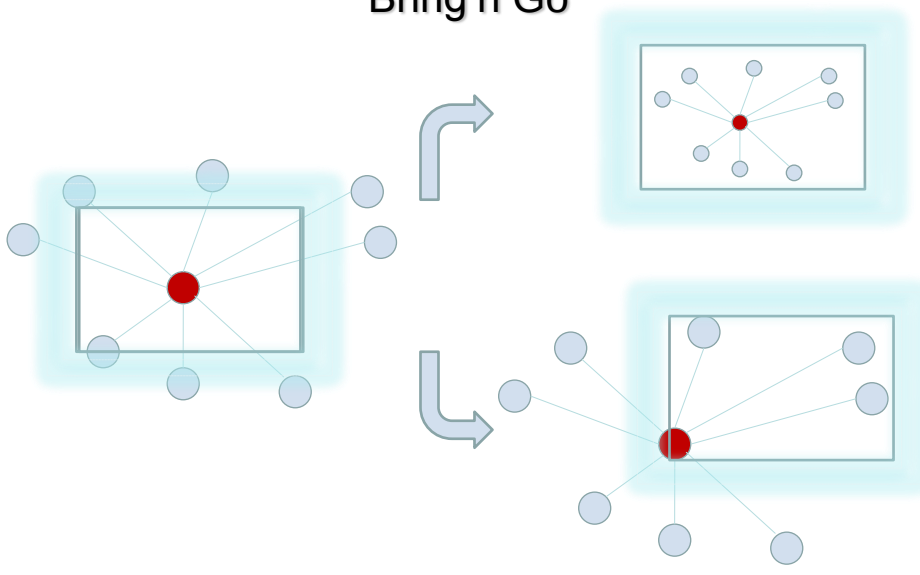
Navigating in large networks

Link Sliding



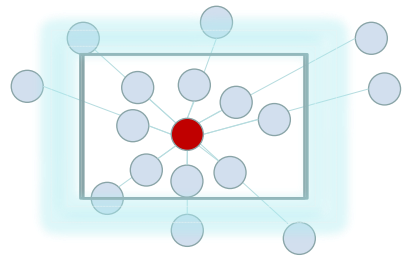
Navigating in large networks

Bring'n Go

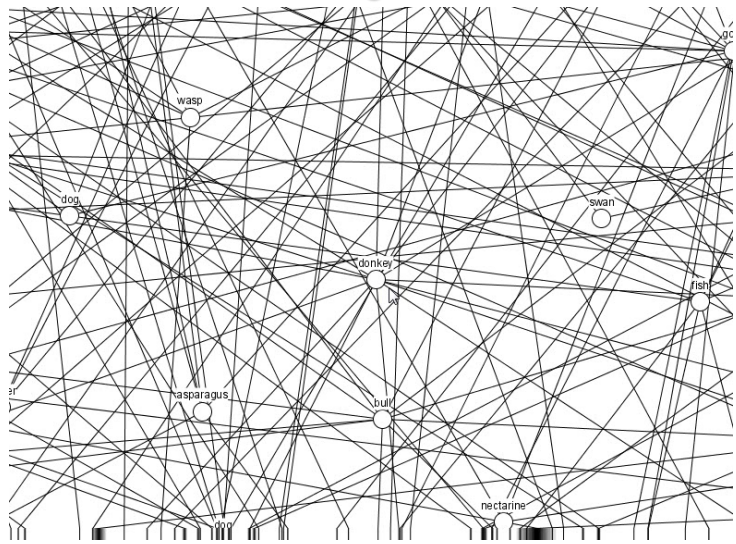


Navigating in large networks
Bring'n Go

- Bring the neighbors into the view

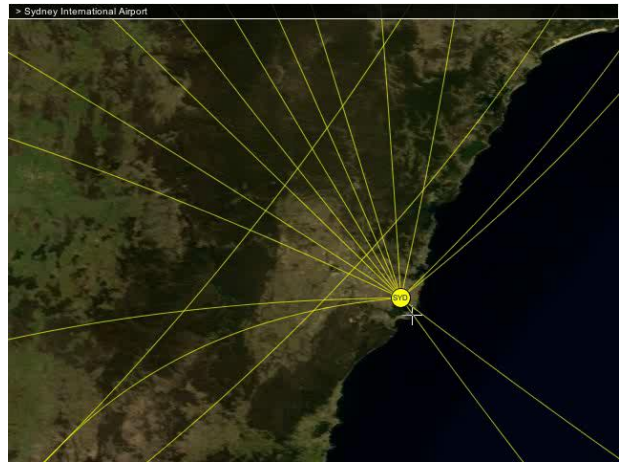


Navigating in large networks
Bring'n Go

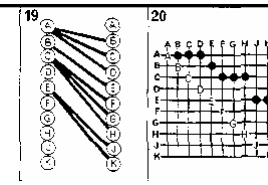


Navigating in large networks

Bring'n Go



Diagrams and Matrices



Directed network

Adjacency matrix

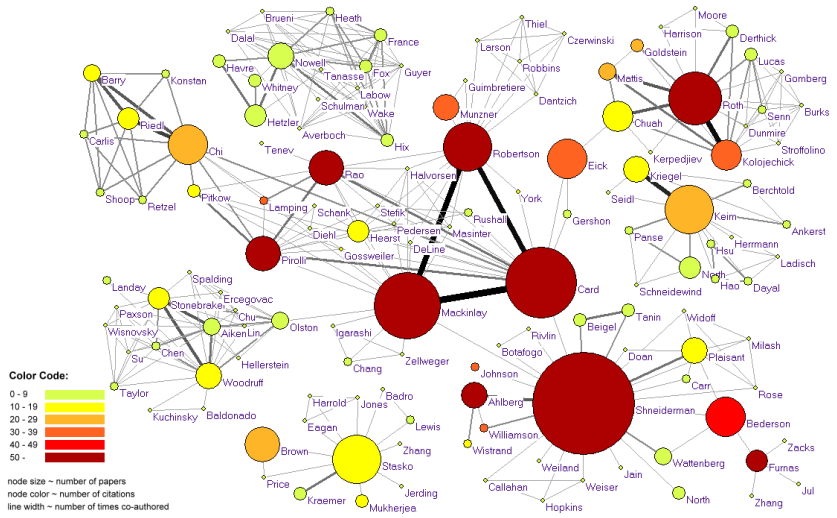
	A	B	C	D
A	0	1	0	0
B	0	0	1	0
C	0	0	0	1
D	0	0	0	0

Destination/target

Source

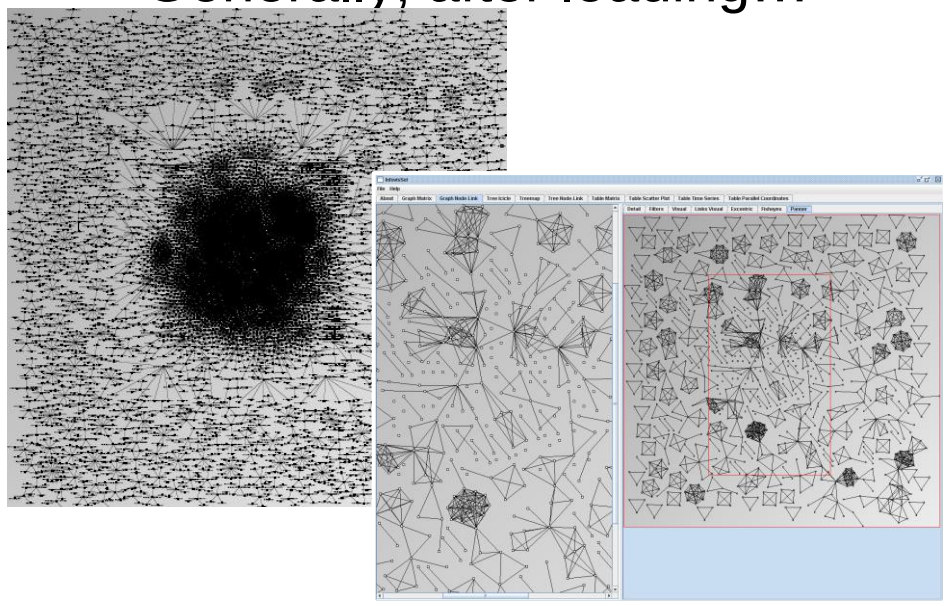
34

InfoVis Co-authoring (K. Börner et al.)



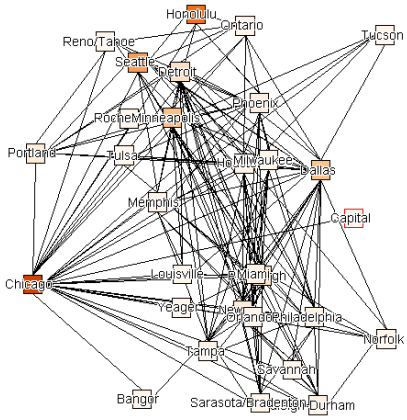
35

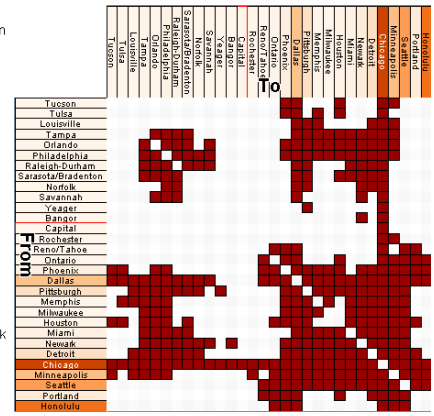
Generally, after loading…



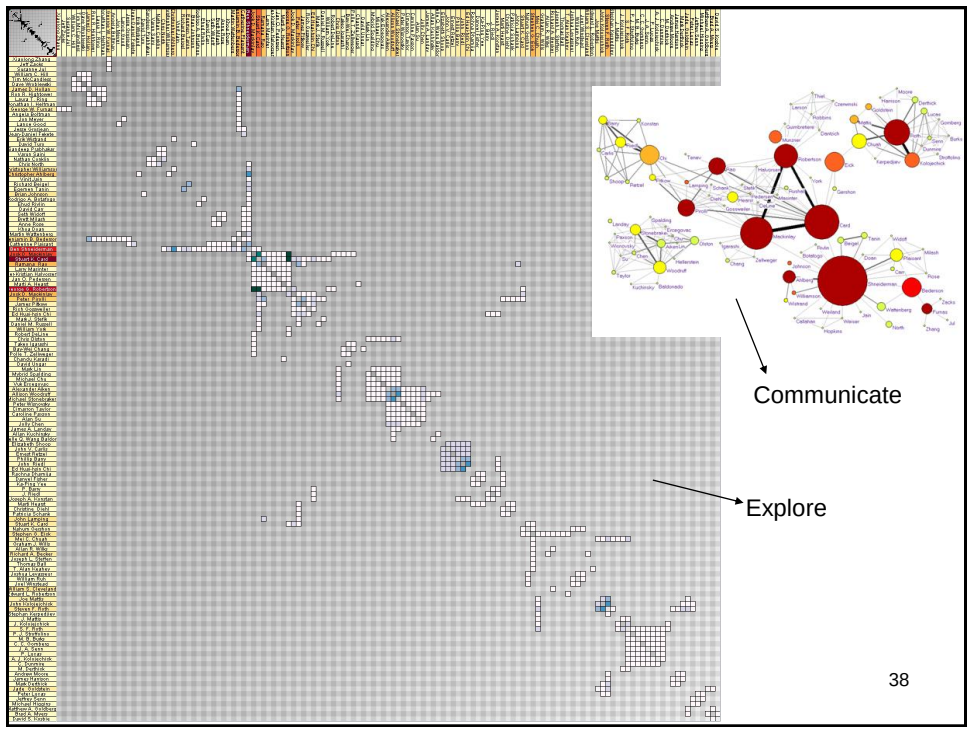
Two classical representations

- Node-Link
- Adjacency Matrix

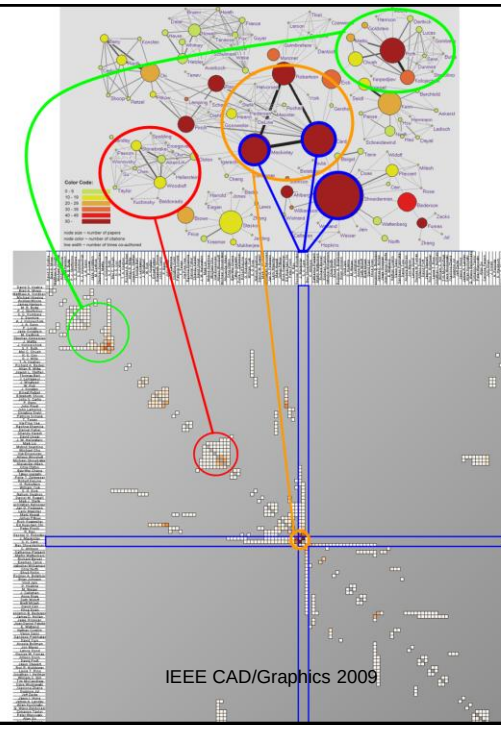




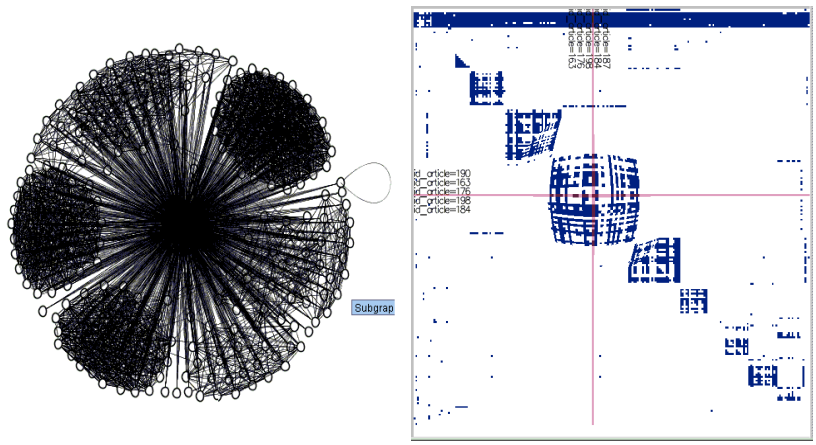
37



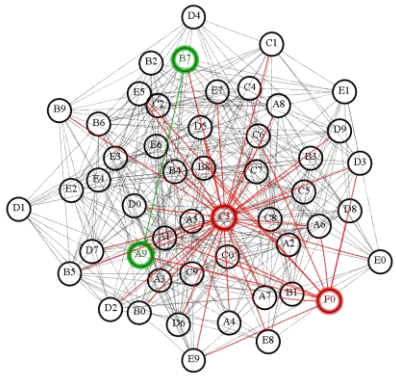
38



Web Site Example



Readability Experiment



41

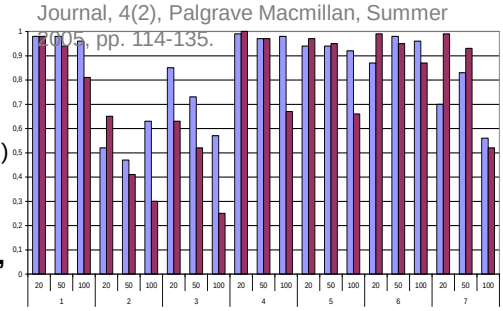
Controlled Experiment: Node Link Diagrams vs. Adjacency Matrices

- **The Tasks:**
- Tasks related to the overview
  - Number of vertices
  - Number of arcs
- Tasks related to graph elements
  - Finding an element (a vertex, a link)
  - Finding the most connected vertex (a central actor, a pivot, a hub)
  - Finding a common neighbor
  - Finding a path
- Random graphs (3 sizes et 3 densities)
- 2 representations: Node-Link + Matrix
- **Results:**
- Node-link diagrams are preferable for small sparse graphs (20 vertices)

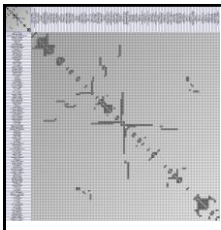
Matrices are more readable *wrt* dense graphs and medium/large graphs (> 20 vertices) *wrt* the selected tasks, except path finding

References:

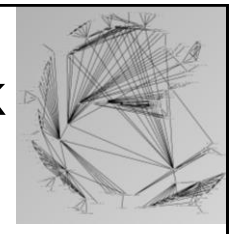
Mohammad Ghoniem, Jean-Daniel Fekete and Philippe Castagliola *Readability of Graphs Using Node-Link and Matrix-Based Representations: Controlled Experiment and Statistical Analysis*, Information Visualization Journal, 4(2), Palgrave Macmillan, Summer 2005, pp. 114-135.



Completion time for the 7 tasks, 3 densities and 2 representations (Node-Link in blue, Matrix in red)

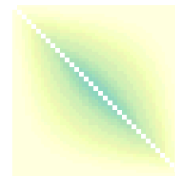


Matrix vs. NodeLink



	Matrix	NodeLink
+	• Usable without reordering • No node overlapping No edge crossing ➔ Readable for dense graphs • Fast navigation • Fast manipulation ➔ Usable interactively • More readable for some tasks	• Familiar • Compact • More readable for path following • More effective for small graphs • More effective for sparse graphs
-	• Less familiar • Use more space • Weak for path following tasks	• Useless without layout • Node overlapping Edge crossing ➔ Not readable for dense graphs • Manipulation requires layout computation

Animation of an activity matrix



MG - NJ - JDF

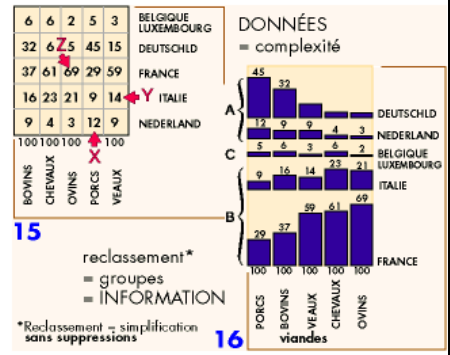
Navigation in an aggregated matrix

45

The Reorderable Matrix

- Introduced by Bertin 67 as a representation for relational data
- Table or Network
- The value table provides details
- The reordered table provides details AND overall structure in the same representation

- Problems:
- how to compute a good ordering?
  - Row and column permutations
- how to assess its quality?



Reordering the Matrix

- Interactive or Automated
- Naïve approach:
  - Define an objective function (e.g. favor diagonal placement and dense clusters)
  - Try all permutations and retain the one that maximizes it
  - Problem : for a n×m table, there are n!×m! configurations
- Four families of methods to reorder a matrix:
  1. Robinsonian
  2. Dimension reduction
  3. Graph linearization methods
  4. Heuristics

Interactive Reordering

- Manual ordering (Bertin's Domino)
- TableLens (Rao & Card 94)
- InfoZoom (Spenke et al. 96)

Hand Reordering with the Domino

- Free one dimension
- Reorder by visual similarity
- Reconnect
- Free the other dimension
- Reorder by visual similarity
- Reconnect
- Take a picture

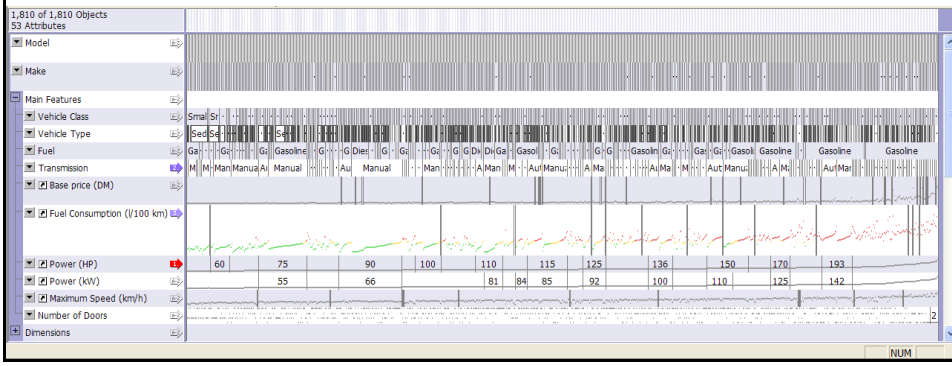
Large dominos

TableLens (Rao & Card 94)

- Interactively reorder based on one attribute
- See the correlation with others
- Fisheye on focus regions
- Only reorder columns, not rows

InfoZoom (Spenke at al. 96)

- Interactively reorder based on one attribute
- See the correlation with others
- Collapse identical values
- Only reorder attributes on rows, not columns



Reordering Tables

Naïve approach:

- Define an objective function (e.g. favor diagonal placement and dense clusters)
- Try all permutations and retain the one that maximizes it
- Problem : for a n×m table, there are n!×m! configurations
- Four families of methods:

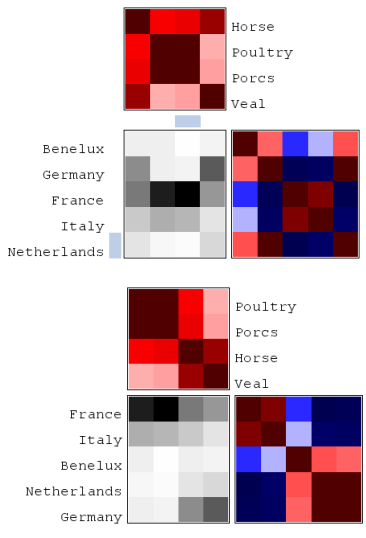
1. Robinsonian
2. Dimension reduction
3. Heuristics
4. Block modeling

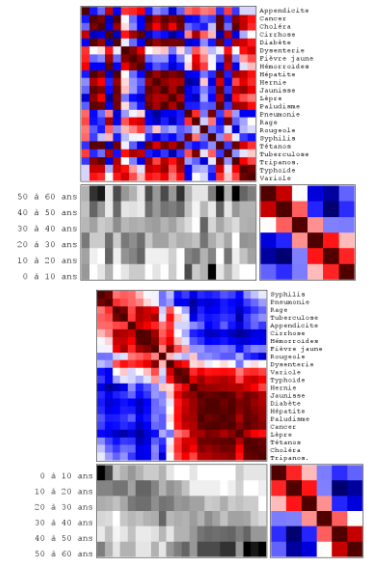
Reordering Tables: Robinsonian (Chun-Houh Chen et al. 2004)

- Goal: Reorder the matrix so that similar rows (resp. columns) are close and dissimilar rows (resp. columns) are farther away

- Load numeric table
- Condition row/columns if necessary
- Choose a distance/dissimilarity for rows & columns
- Compute the distance matrix for rows & for columns
- Compute the row (resp. column) permutation to get a Robinsonian distance matrix
- Apply the permutation to the table row (resp. column)

GAP:http://gap.stat.sinica.edu.tw/Software





Reordering Tables: Conditioning

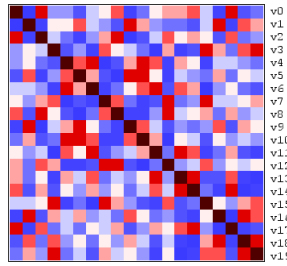
- Standard operation on tables
- Some tables need conditioning
  - For example, the CIA World Fact Book
    - Area (sq km)
    - Birth rate(births/1000 population)
    - Death rate (deaths/1000 population)
- Values vary widely in range
- Conditioning make them comparable:
  - e.g. center mean to 1 and variance to 1
- Other tables don't need it
  - For example student grade per class (all between 0 and 5 or 20)

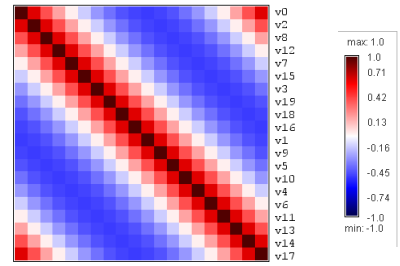
Reordering Tables: Distance / dissimilarity

- Compute row/row and column/column distance
- Several popular distances between row x and row y
- L0 :
- L1 : $\sum|x_i - y_i|$
- L∞: $\sqrt{\sum(x_i - y_i)^2}$
- Pearson's correlation $\max|x_i - y_i|$
- $$\frac{\sum_k (d_{ik} - \bar{d}_{i\cdot})(d_{jk} - \bar{d}_{j\cdot})}{\sqrt{\sum_k (d_{ik} - \bar{d}_{i\cdot})^2}\sqrt{\sum_k (d_{jk} - \bar{d}_{j\cdot})^2}},$$
- Complexity: $\max(n^2, m^2)$

Reordering Tables: Robinsonian Matrix

- The distance matrix is positive and symmetric
- In a perfect world, we would like to find a consistent similarity matrix:
- maximum on the diagonal and decreases off the diagonal
- This is called a Robinsonian Matrix (R-matrix)
- A pre-R matrix is a similarity matrix that can be permuted to become an R-matrix
- If that permutation exists, we want to find it
- Otherwise, we want a good approximation





Reordering Tables: R-matrix computation

- Known methods to solve the simple case:

1. Atkins et al. 98 describe a “Spectral” method that solves the problem when the R-matrix exists
2. TSP direct method

- Chun-Houh Chen uses two methods

3. Hierarchical clustering + seriation
   Several options for clustering
4. His own Ellipse seriation technique

A SPECTRAL ALGORITHM FOR SERIATION AND THE CONSECUTIVE ONES PROBLEM

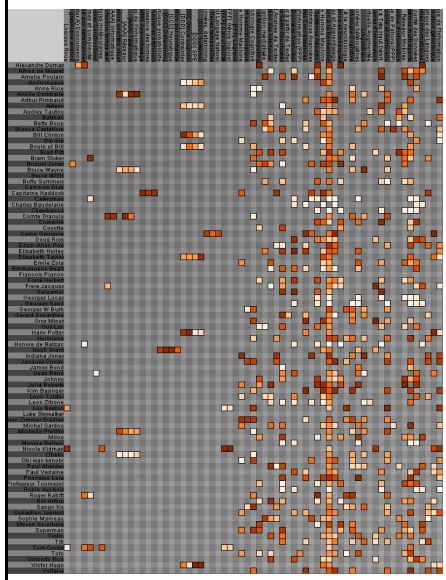
JONATHAN E. ATKINS, ERIK G. BOMAN, AND BRUCE HENDRICKSON (SIAM 98)

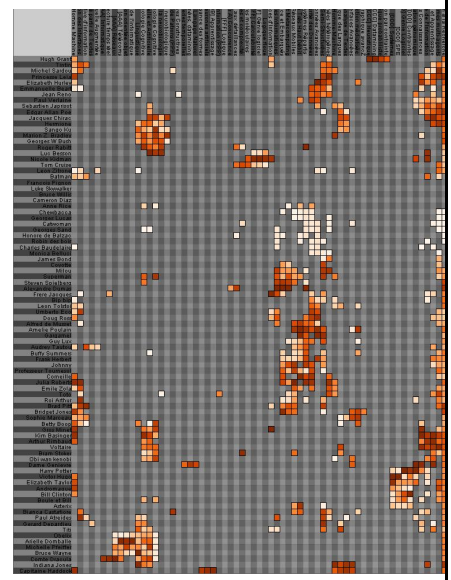
- A distance matrix $A_{i=1..n,j=1..n}=[a_{i,j}]$
- A is positive and symmetric: $a_{i,j} \geq 0$ and $a_{i,j} = a_{j,i}$
- Computing the permutation π to transform a pre-R matrix into A
- BEGIN MAGIC
- π can be computed from the eigenvector x with the smallest non-null eigenvalue of the Laplacian of A: it is the permutation that orders the components of *x*.
- END MAGIC
- An Eigenvector *x* of a matrix square matrix M of size *n* is a vector
- $x=[x_1, x_2, \ldots, x_n]^T$ such that
- $Mx = \lambda x$ where λ is a real number called the Eigenvalue of *x*
- M has *n* eigenvalues/eigenvectors that can be sorted by eigenvalue
- The Laplacian of A, $L_A = D_A - A$ where $D_A=[d_{i,i}]$, $d_{i,i} = \sum_{i=1..n} a_{i,j}$
- Computing this Eigenvector is sub-quadratic in practice

Reordering Table: TSP

- We can compute a seriation by solving the Traveling Salesman Problem (TSP).
- TSP is NP complete but heuristics running in polynomial time exist (Lin&Kernighan 73) and free programs are available:
- LKH http://www.akira.ruc.dk/~keld/research/LKH/
- Concorde http://www.tsp.gatech.edu/concorde.html
- TSP has been used for matrix seriation
- 1 page article (Lenstra 74) “It performs well…”
- (Climer&Zhang 04) applied it to microarray data
- (Henry&Fekete 06) applied it to tables and graphs

TSP Reordering



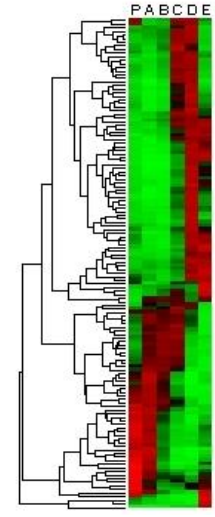


Hierarchical clustering + seriation

- Method popular in Bioinformatics: Eisen et al. 98

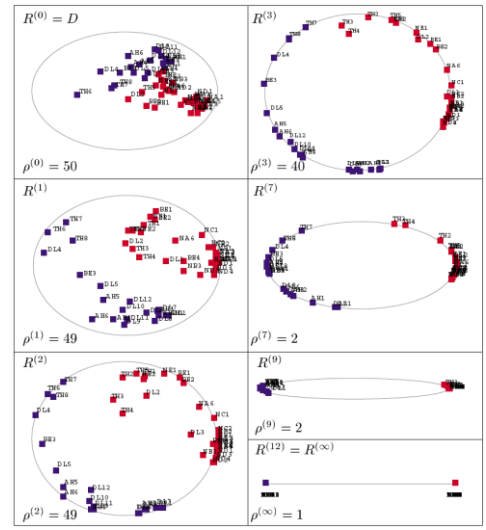
1. Hierarchical clustering from distance matrix
2. Linear order constrained by the binary tree

- For n leaves, 2^{n-1} linear orderings possible
- We want an order that minimizes the sum of distances between consecutive items (TSP)
- (Bar-Joseph et al. 02) describe a $O(n^3)$ algorithm
- (Brandes 07) describes a $O(n^2 \log(n))$ algorithm
- (Deĭneko&Tiskin 07) describe a complex $O(n^2)$ algorithm



Ellipse seriation

- Method based on computing the correlation matrix D, the distance matrix using Pearson's correlation coefficient
- Iteratively, a second correlation matrix is computed from that
- $\varphi(A)=D$, $R^1=\varphi(D)$, $R^{n+1}=\varphi(R^n)$
- With no theoretical reason, when looking at the projection of the columns of R^n onto the plane of the 2 main eigenvectors, we have an ellipse when $n\to\infty$
- Unrolling the ellipse, we have a good approximation of an R-matrix



Results

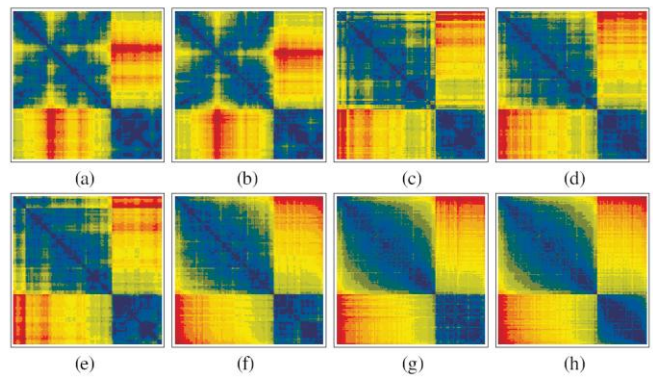


Figure 4: Permuted Euclidean distance maps for Iris data with eight seriation algorithms: (a) farthest insertion spanning; (b) nearest insertion spanning; (c) single linkage tree; (d) complete linkage tree; (e) average linkage tree; (f) GAP rank-one tree; (g) GAP rank-two ellipse; (h) GAP double ellipse.

Reordering Tables: Dimension reduction

- (Chauchat&Risson 95) use 2 dimension reduction techniques to reorder matrices:

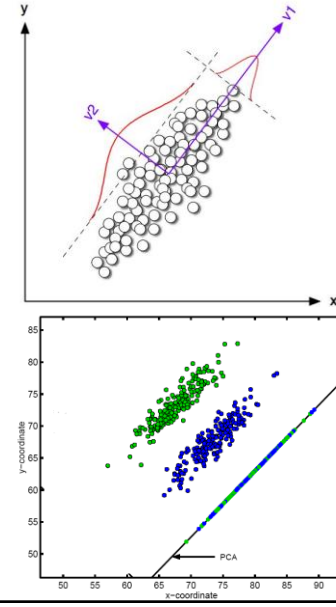
1. Principal Component Analysis (PCA)
2. Correspondence Analysis (CA)

- Their methods are available in the AMADO system

- Both method assume linear relationships between rows and columns

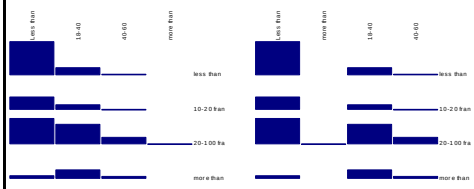
Principal Component Analysis

- Each column is considered as a vector in high-dimensional space
- All the vectors define a cloud extending in *n* dimension
- PCA computes the principal axes of this cloud
- The first component is the axis where the vectors are most spread
- The second component is orthogonal to the first and is the second axis where the vectors are most spread
- Etc.
- If the vectors are linearly correlated, they are aligned on the principal axis
- Otherwise, they spread on several principal axis
- The components are computed from A by centering the columns to their mean values and computing the eigenvectors/eigenvalues
- The matrix order is the order of the coordinates in the first eigenvector
- Time is bounded by the Eigenvector computation
- Sub-quadratic in practice
- Compute the first Eigenvector twice (Rows and Columns)



Correspondence Analysis

- Designed by (Benzécri 74) for contingency tables
- Compute vectors as difference from the expected values
- Then, same as PCA except compute Eigenvector ONCE
- Very good results for contingency tables!

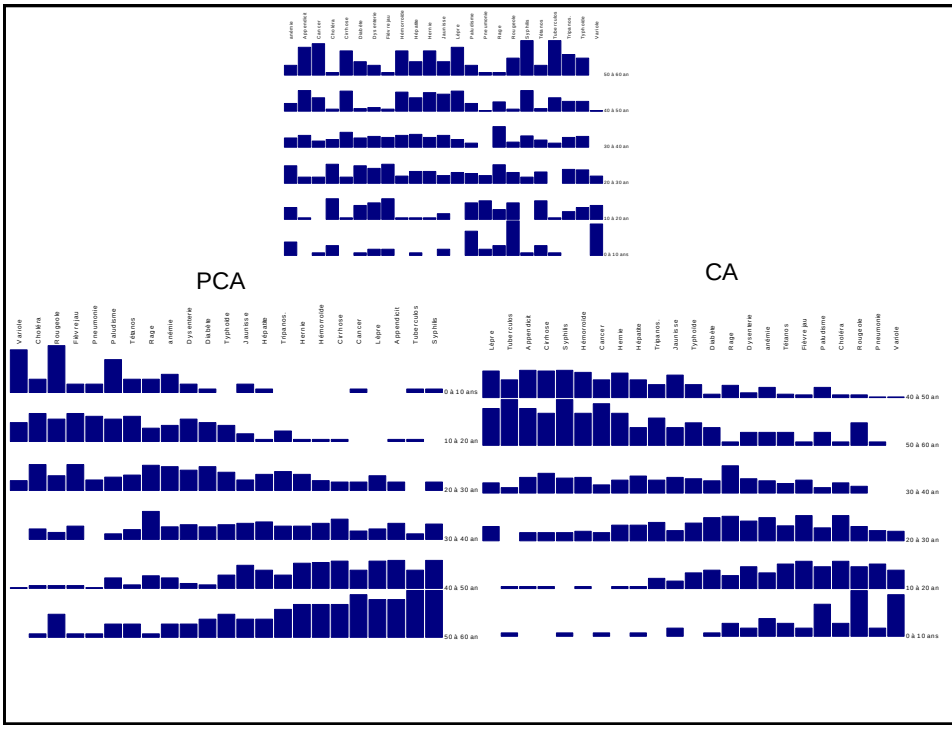


Thefts Value/Age	Less than 18	18-40	40-60	more than 60	**Row total**	Row percent
less than 10 francs	113	27	5	0	145	32,01%
10 ˆ 20 francs	44	19	4	1	68	15,01%
20 ˆ 100 francs	89	68	26	3	186	41,06%
more than 100 francs	12	32	10	0	54	11,92%
Column total	258	146	45	4	453	100,00 %
Column percent	56,95%	32,23%	9,93%	0,88%	100,00 %	

Expected percentages	18,23%	10,32%	3,18%	0,28%		32,01%
Col%*Row%	8,55%	4,84%	1,49%	0,13%		15,01%
	23,38%	13,23%	4,08%	0,36%		41,06%
	6,79%	3,84%	1,18%	0,11%		11,92%
Column percent	56,95%	32,23%	9,93%	0,88%		

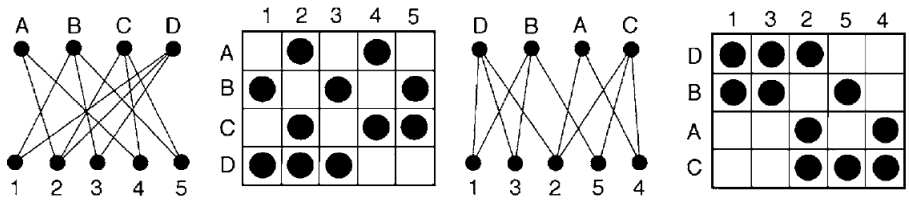
Expected value	82,582781	46,732892	14,403974	1,280332
Expected%*total	38,728477	21,916115	6,754969	0,600441
	105,93377	59,94702	18,476821	1,642341
	30,754967	17,403974	5,364284	0,476812

Difference in %	6,71%	-4,36%	-2,08%	-0,28%
	1,16%	-0,64%	-0,61%	0,09%
	-3,74%	1,78%	1,66%	0,30%
	-4,14%	3,22%	1,02%	-0,11%



Reordering Tables: Heuristics

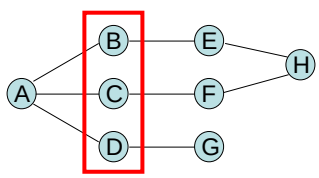
- (Siirtola&Mäkinen 05) Consider a Table as a bipartite graph
- Then use a heuristics to minimize the number of crossings
- find the best permutation of rows and columns
- Proved to be a NP-complete problem
- Barycenter Heuristics repeatedly reorder the vertices according to the average of their adjacent vertices in the opposite vertex set.



Mixed approach

[Henry and Fekete, IHM'06] [Henry and Fekete, InfoVis'06]

- Place actors with similar connection patterns next to each other

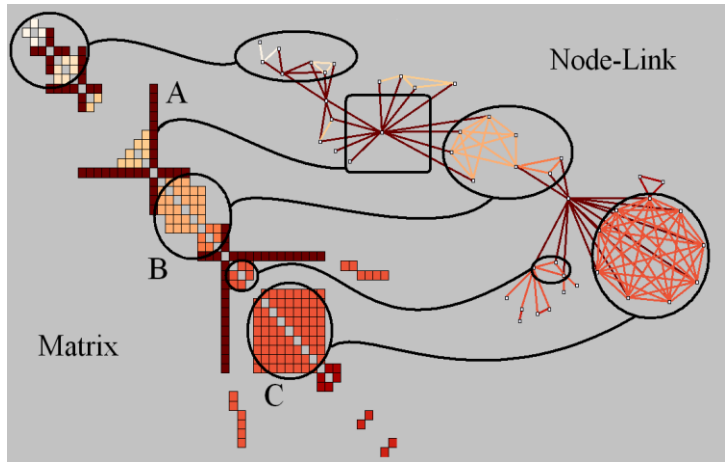


	A	B	C	D	E	F	G	H
A	0	1	1	1	0	0	0	0
B	1	0	0	0	1	0	0	0
C	1	0	0	0	0	1	0	0
D	1	0	0	0	0	0	1	0
E	0	1	0	0	0	0	0	1
F	0	0	1	0	0	0	0	1
G	0	0	0	1	0	0	0	0
H	0	0	0	0	1	1	0	0

	A	B	C	D	E	F	G	H
A	0	1	1	1	2	2	2	3
B	1	0	2	2	1	3	3	2
C	1	2	0	2	3	1	3	2
D	1	2	2	0	3	3	1	4
E	2	1	3	3	0	2	4	1
F	2	3	1	3	2	0	4	1
G	2	3	3	1	4	4	0	5
H	3	2	2	4	1	1	5	0

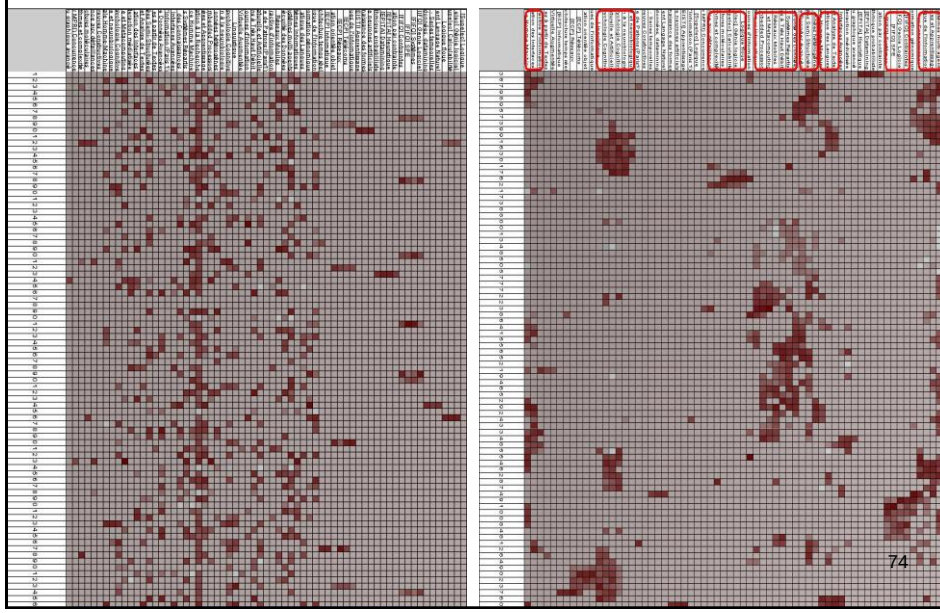
→ Add information to the adjacency matrix

Visual Patterns with Ordered Matrices



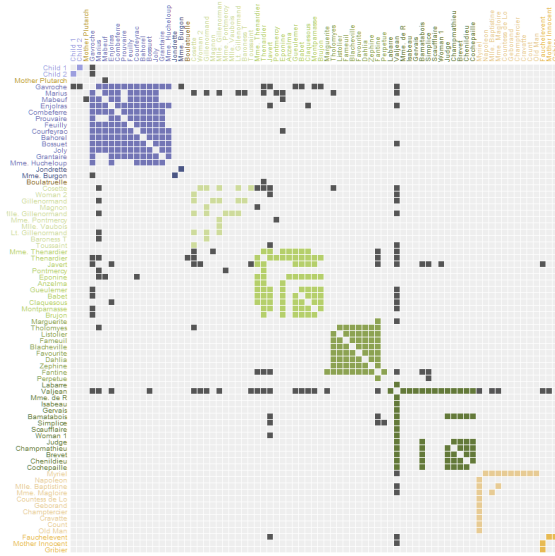
73

Ordonner une matrice



74

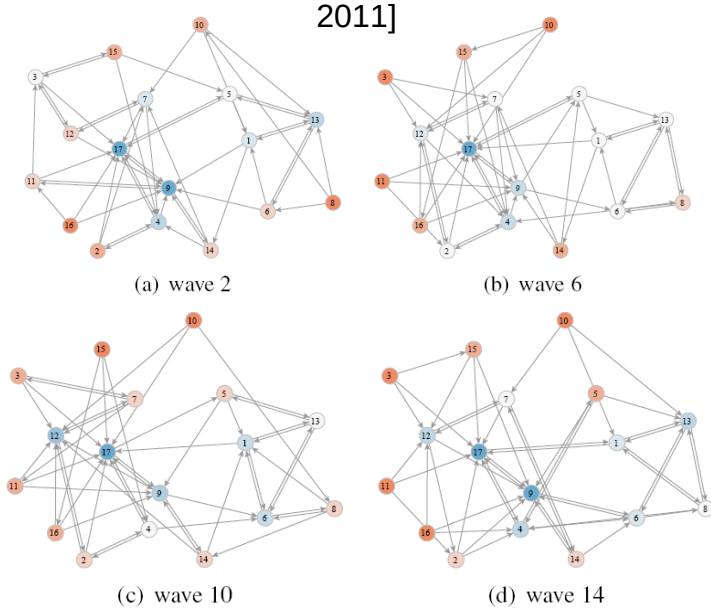
Les personnages dans Les Misérables



75

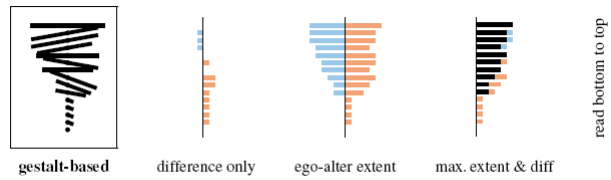
Longitudinal Social Networks

Ulrik Brandes and Bobo Nick, Univ. Konstanz [InfoVis 2011]



(a) wave 2 (b) wave 6

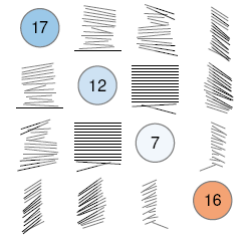
(c) wave 10 (d) wave 14



data: (1,1),(1,2),(1,2),(1,2),(1,2),(3,4),(3,5),(4,6),(5,5),(5,6),(6,6),(6,5),(7,6),(7,6),(7,7)

While 2 befriends popular 17 only late, 16 17 seems to stand no chance at that. In fact, 16 is unsuccessful in entering the tightly knit group of 17, 12, and 7.

The relationships between 7 14, 3 15, or 10 15 are rather more complicated. The latter is no surprise as 10 is fairly unpopular with many.

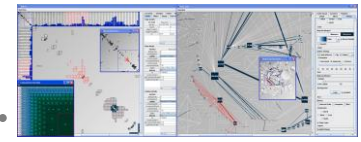


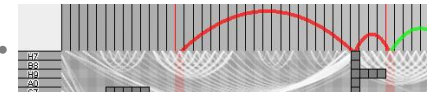
Gestaltmatrix of Newcomb Fraternity Data, showing the evolution using all 4080 data points (rankings) from 15 waves. Labels on the diagonal are numbered according to Nordlie [42] and colored according to the standardized overall popularity of the corresponding actor. The sorting of actors is according to a spectral approach.

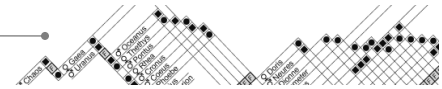
Breakthrough in Social Network Visualization: Improving Matrices

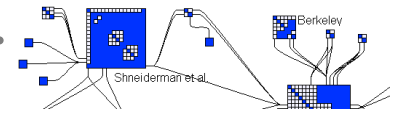
Several representations:

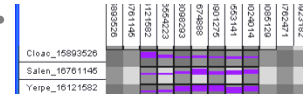
1. **Combined**
   - MatrixExplorer (Henry&Fekete InfoVis'06)
2. **Augmented**
   - MatLink (Henry&Fekete Interact'07, **Best Paper**)
   - GeneaQuilts (Bezerianos et al. InfoVis'10)
3. **Hybrid**
   - NodeTrix (Henry et al. InfoVis'07)
   - CoCoNutTrix (Isenberg et al. CG&A'09)
4. **Multiscale**
   - ZAME (Elmqvist et al. PacificVis'08)





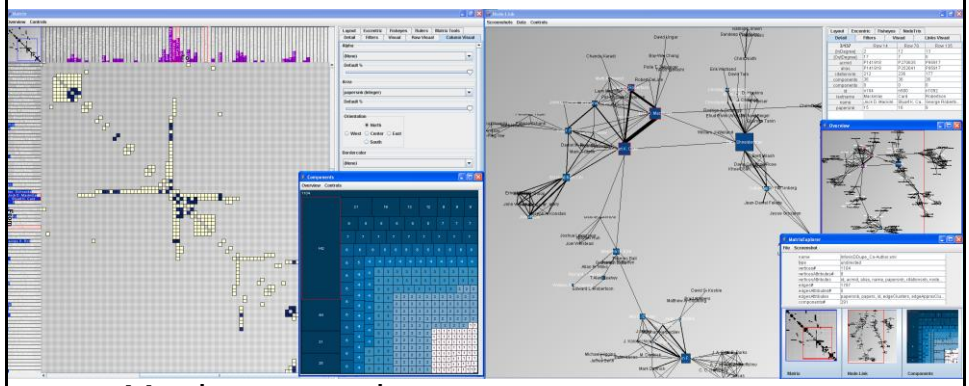






MatrixExplorer [Henry&Fekete06]

Combined representation

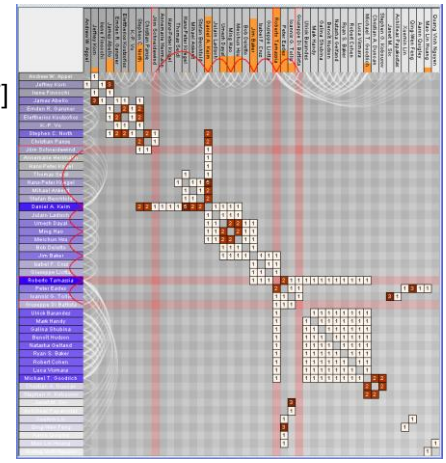


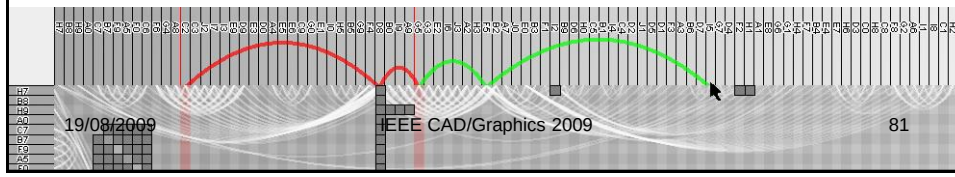
- Matrices to explore
- Node-Link diagrams to present findings

MatLink[Henry&Fekete07]

Augmented representation

- Augmenting matrices with interactive links
- Solving the path-related tasks problem for matrices





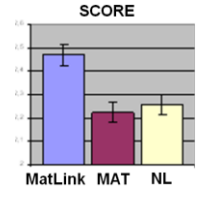
MatLink significantly improves matrices

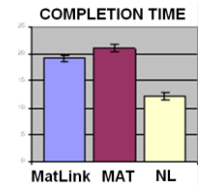
- Controlled experiment
  - 3 vis. x 6 datasets x 5 tasks

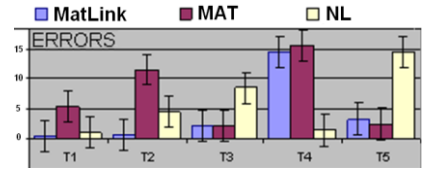
  Matrix , Node-Link, MatLink

  Data: From almost-trees
  To complete-graphs
  Including small-world networks

  Tasks:
  1. CommonNeighbour,
  2. ShortestPath,
  3. MostConnected,
  4. ArticulationPoint,
  5. LargestClique



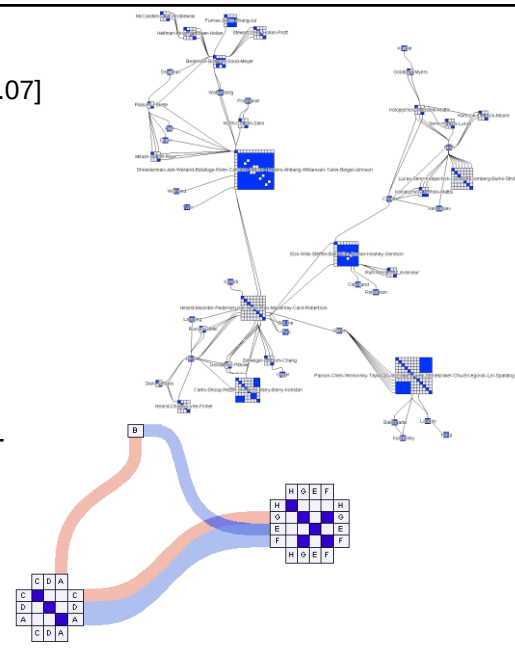




NodeTrix[Henry et al.07]

Hybrid representation

- Designed for small-world networks
  - Globally sparse
  - Locally dense
- Visualizing dense sub-graphs as matrices
- Interact to create, edit and remove the matrices

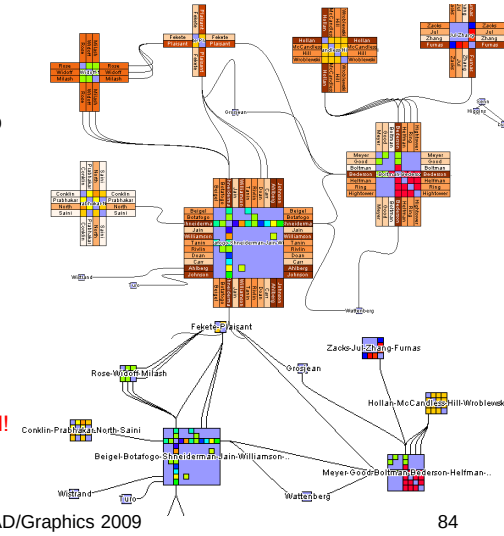


NodeTrix: the NetVis Nirvana?

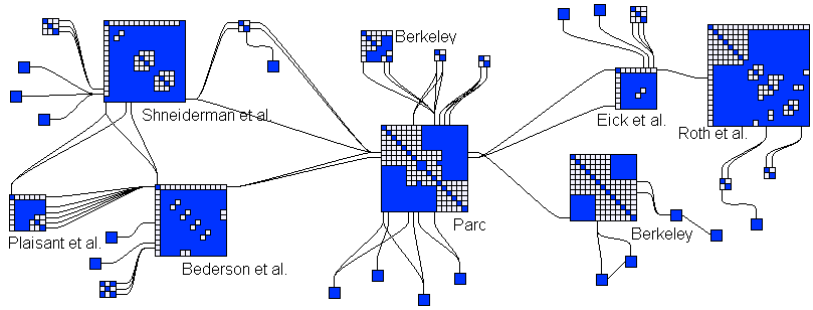
- ✓ Can you see every node?
- ✓ Can you count each node's degree?
- ✓ Can follow every link from its source to its destination?
- ✓ Can you identify clusters and outliers?

- Node Labels
- Link Labels (excentric labels?!)
- … even cluster labels
- Node Attributes
- Link Attributes
- … even clusters attributes
- Directed Graph (links width?!)

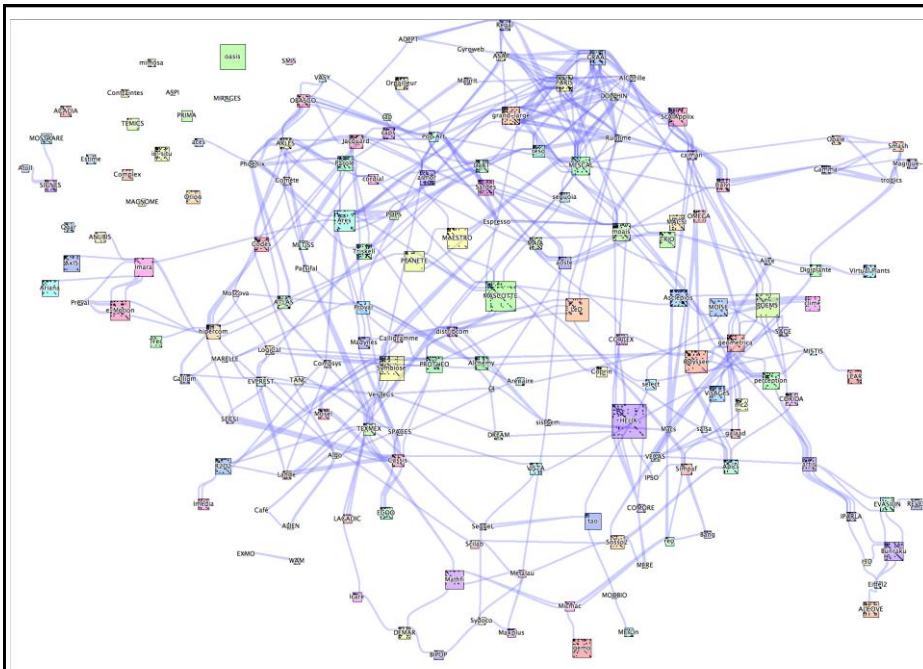
… But… beware the graphics overload!



Visual Patterns



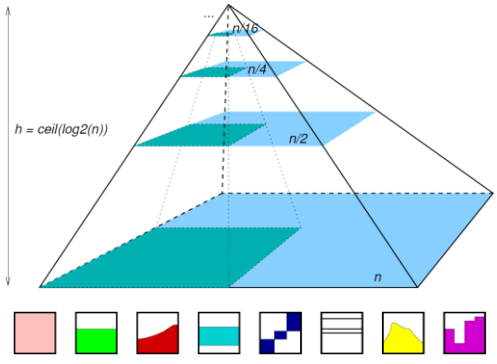
Infovis Coauthorship (133 actors)



ZAME: Interactive Large-Scale Graph Visualization

Visualize very large networks:
- Larger than 10^7 vertices and edges

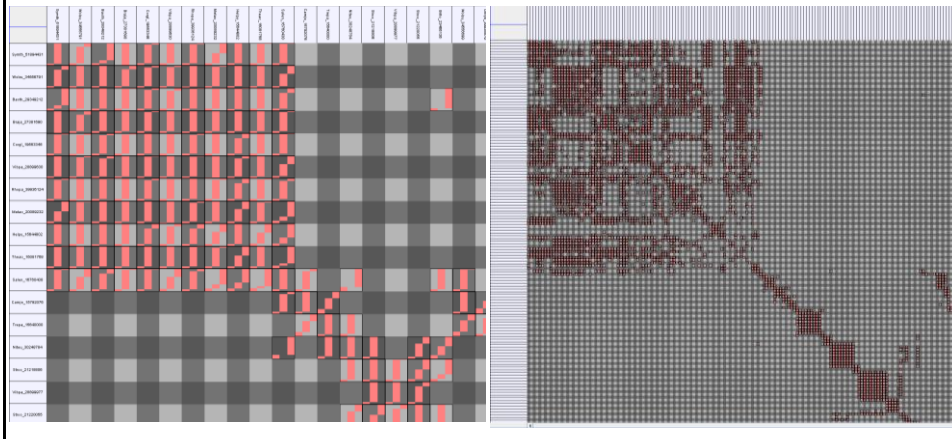
- Reorder
- Create a pyramid
- Aggregate attributes
- Visualize using enhanced glyphs

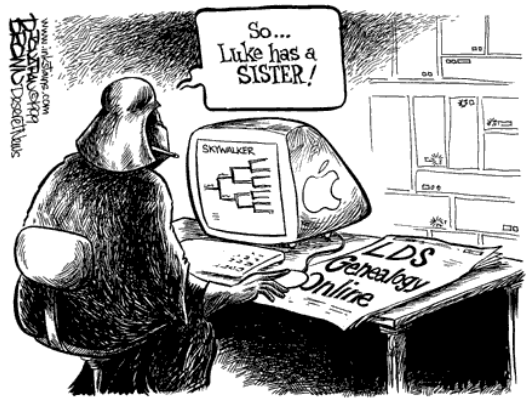


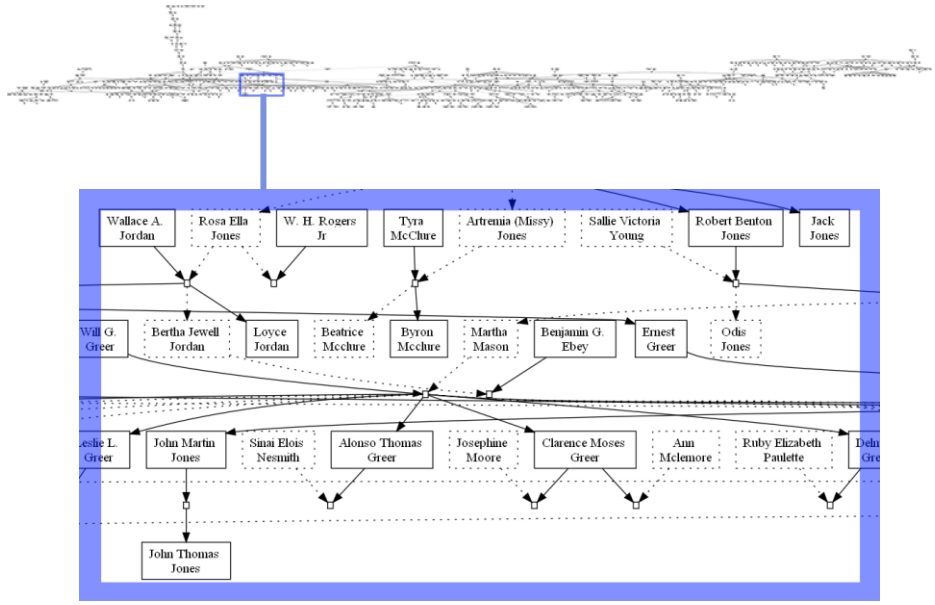
(Elmqvist et al. 08)

Visualization and Navigation

- Space – Scale + Resolution
- Accelerated graphics: Tile Rendering & OpenGL + shaders

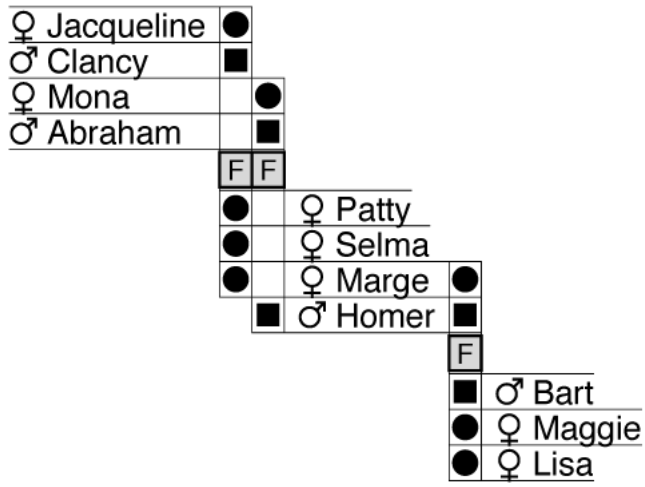



So...
Luke has a
SISTER!
SKYWALKER
LDS
Genealogy
Online


Wallace A. Jordan
Rosa Ella Jones
W. H. Rogers Jr
Tyra McClure
Artremia (Missy) Jones
Sallie Victoria Young
Robert Benton Jones
Jack Jones
Will G. Greer
Bertha Jewell Jordan
Loyce Jordan
Beatrice Mcclure
Byron Mcclure
Martha Mason
Benjamin G. Ebey
Ernest Greer
Odis Jones
John Martin Jones
Sinai Elois Nesmith
Alonso Thomas Greer
Josephine Moore
Clarence Moses Greer
Ann Mclemore
Ruby Elizabeth Paulette
John Thomas Jones

GeneaQuilts

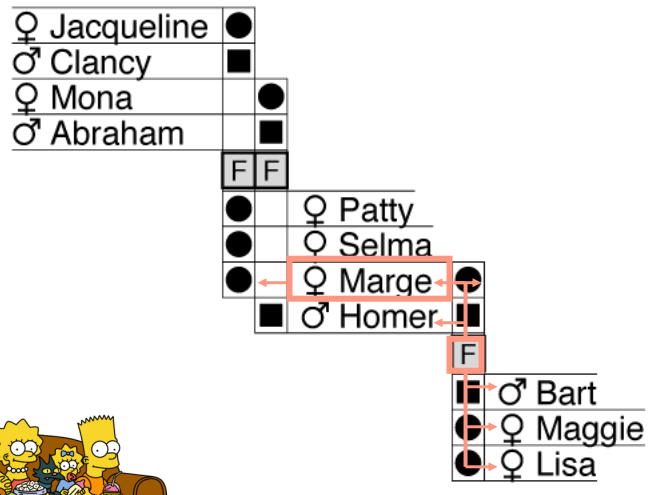
GeneaQuilts teaser

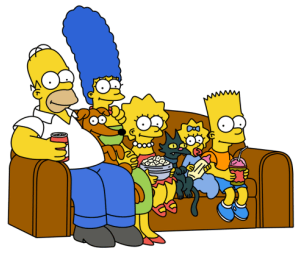


91

GeneaQuilts

GeneaQuilts Vis

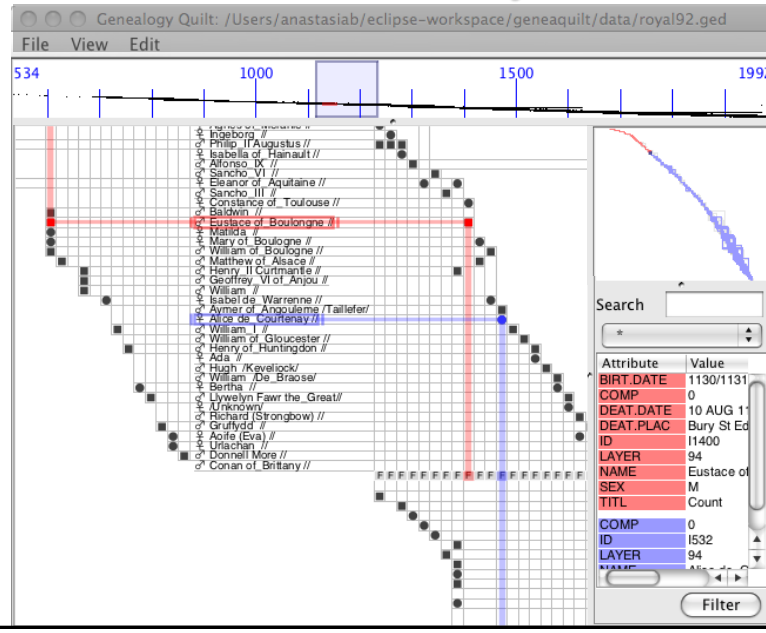




92

GeneaQuilts

GeneaQuilts system



93

GeneaQuilts

Layout algorithm

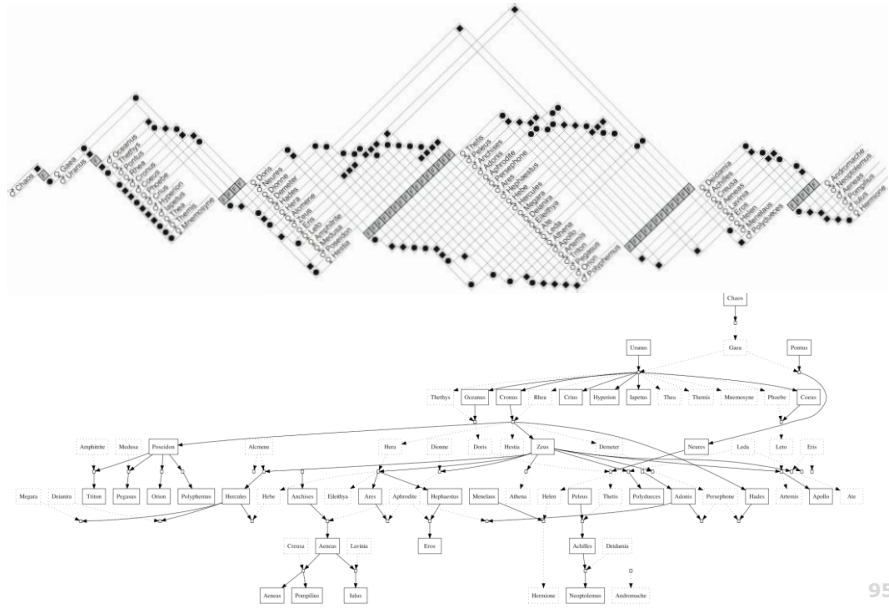
- Generations
  - layer assignment algorithm [Gansner et al.]
- Order in generations
  - minimize link crossings [Siirtola & Mäkinen]
  - ensures siblings and parents close

Use AT&T dot library

94

GeneaQuilts

Layout algorithm



95

Conclusion

- Visualization of Networks has been greatly improved in the last years
- Novel representations are denser and more expressive
  - Though they require a little training
- Huge and dense networks can be visualized
  - Relations between clients, suppliers, employees
  - Aggregated over long periods of time
- Needs more research to understand how reordering leads to better understanding
- Visit our site: http://www.aviz.fr

Challenges

- Exploring Very Large / Very Dense Networks
  - Top-Down methods
  - Flexible clusterings
- Linking Exploration and Modeling
- Reordering Methods for Matrices
  - What is a good order and why?
  - Orderings for directed graphs
  - Multi-Scale Ordering (top-down methods)

References

- M. Ghoniem, J-D. Fekete and P. Castagliola. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis 04), pages 17-24, Austin, TX, October 2004. IEEE Press.
- F. van Ham, Using Multilevel Call Matrices in Large Software Projects Proc. IEEE Symp. Information Visualization 2003, IEEE CS Press, pp.227-232, 2003.
- J. Abello and F. van Ham, Matrix Zoom: A Visual Interface to Semi-External Graphs Proc. IEEE Symp. Information Visualization 2004, IEEE CS Press, pp. 183-190, 2004.
- Harri Siirtola and Erkki Mäkinen. Constructing and reconstructing the reorderable matrix.. Information Visualization, 4(1):32-48, 2005.
- N. Henry and J-D. Fekete. MatrixExplorer: a Dual-Representation System to Explore Social Networks. IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2006), 12(5):8 pages, to be published, September-October 2006.
- N. Henry and J-D. Fekete. MatLink: Enhanced Matrix Visualization for Analyzing Social Networks. In C. Baranauskas, P. Palanque, J. Abascal and S. D. J. Barbosa, editors, Human-Computer Interaction – INTERACT 2007, volume 4663 of LNCS, pages 288–302, 2007. Springer.
- N. Henry, J-D. Fekete and M. J. McGuffin. NodeTrix: a Hybrid Visualization of Social Networks. IEEE Transactions on Visualization and Computer Graphics, 13(6):1302-1309, 2007.
- Christopher Mueller, Benjamin Martin, and Andrew Lumsdaine. A comparison of vertex ordering algorithms for large graph visualization. Visualization, 2007. APVIS '07. 2007 6th International Asia-Pacific Symposium on, pages 141-148, 5-7 Feb. 2007.
- N. Elmqvist, T-N. Do, H. Goodell, N. Henry and J-D. Fekete. ZAME: Interactive Large-Scale Graph Visualization. In Proceedings of the IEEE Pacific Visualization Symposium 2008, pages 215–222, March 2008. IEEE Press.
- A. Bezerianos, P. Dragicevic, J-D. Fekete, J. Bae and B. Watson. GeneaQuilts: A System for Exploring Large Genealogies. IEEE Transactions on Visualization and Computer Graphics, 16(6):1073–1081, Nov-Dec 2010.

Free Software

- Java
  - Prefuse, the InfoVis Toolkit, Improvise, JUNG, Guess, Gephi
- C++
  - Tulip
- Web
  - Flare (flash)
  - Protovis, Javascript InfoVis Toolkit (JIT), D3

99