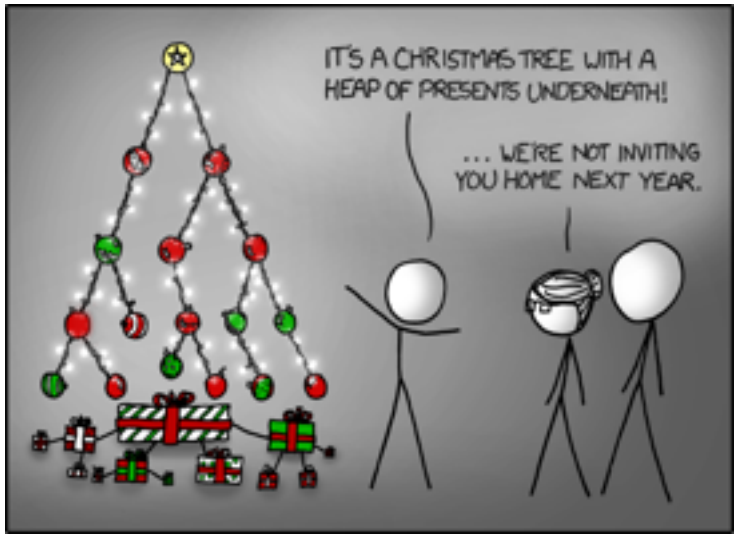
CS109 Data Science
Trees, Networks & Databases

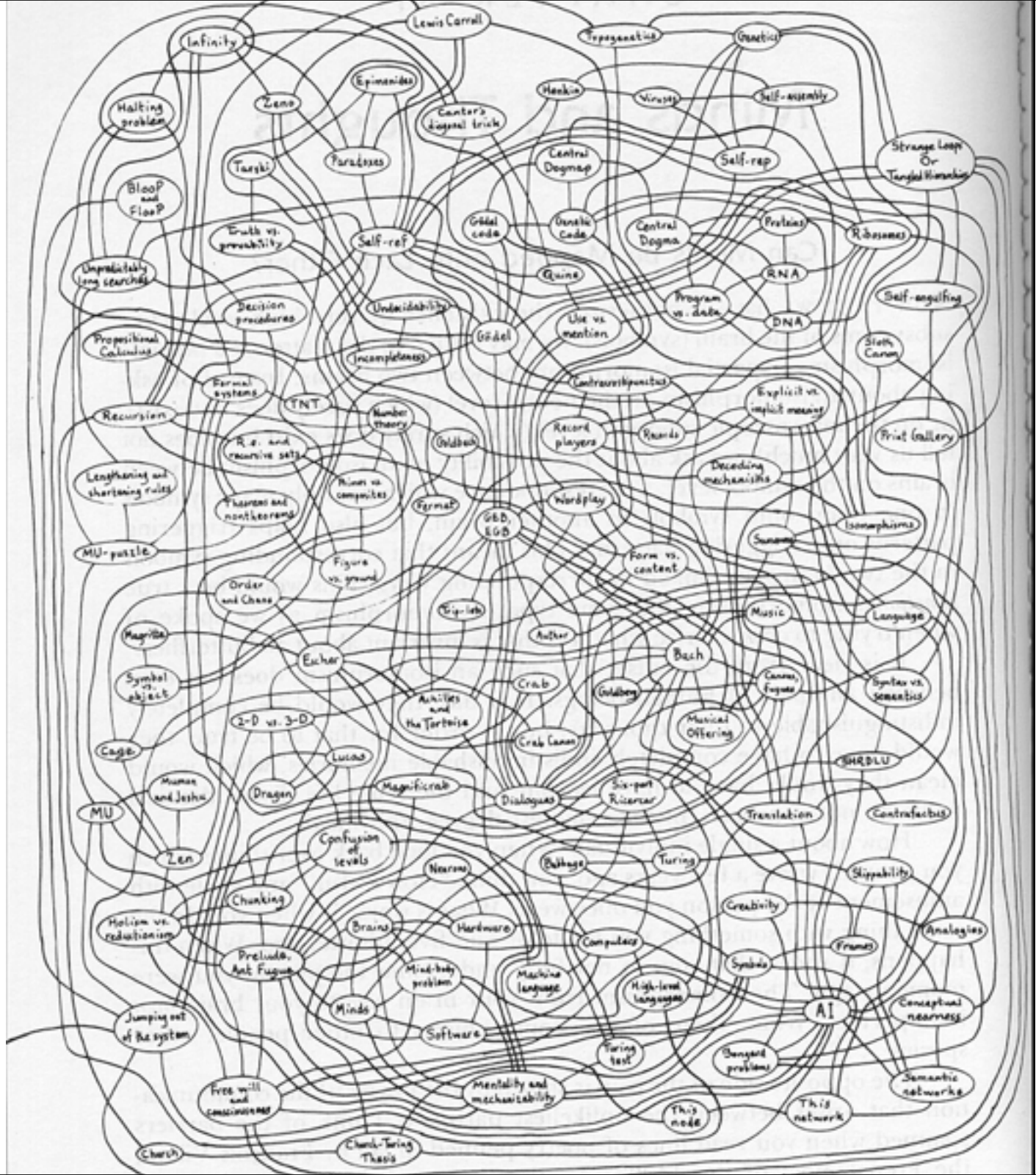
Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu

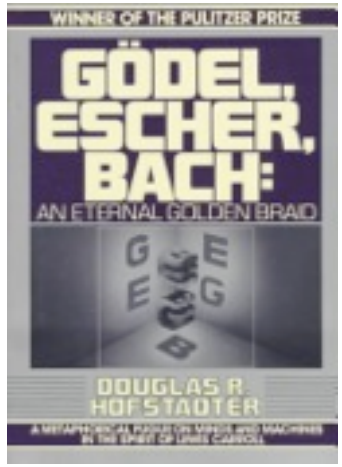


xkcd

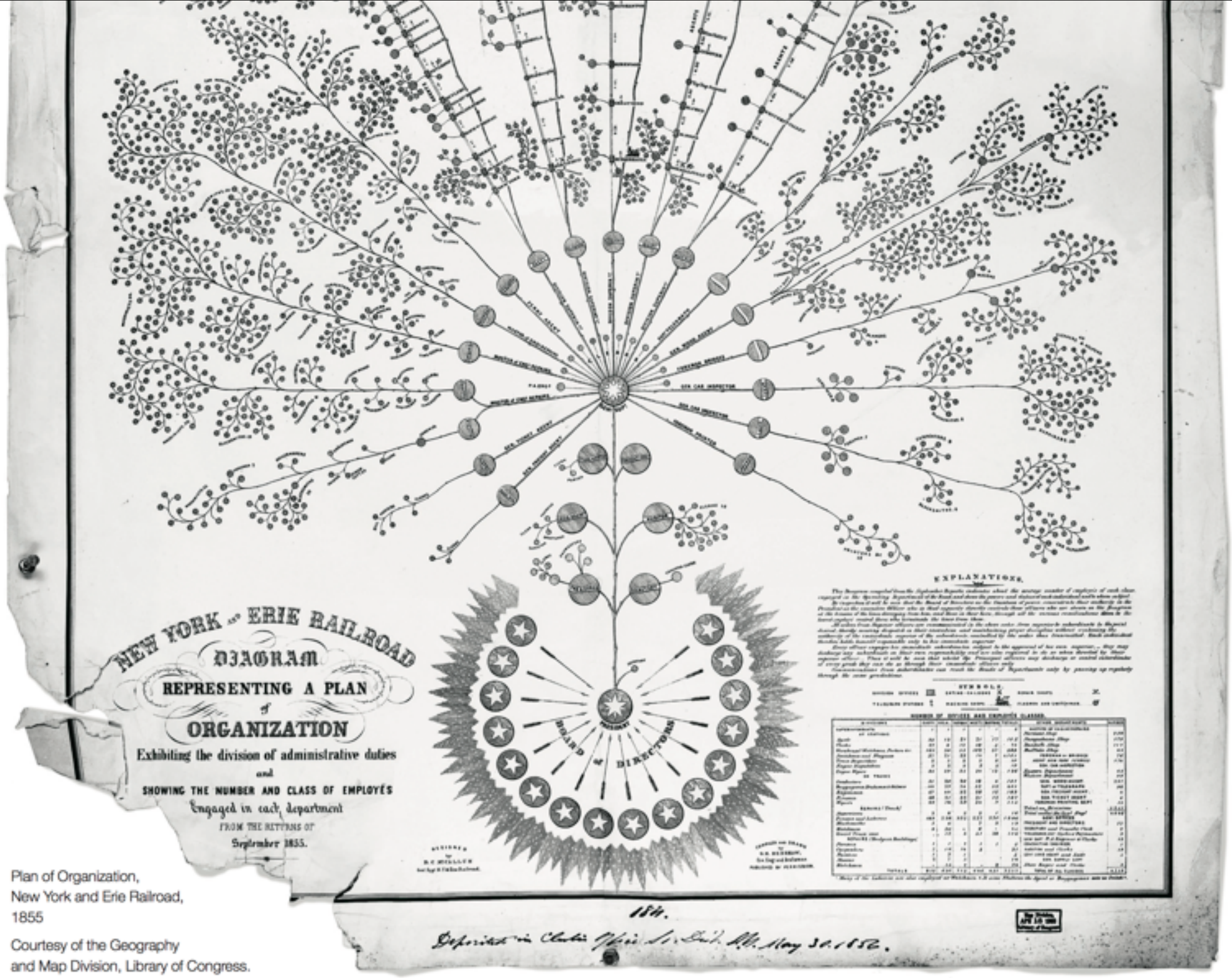
This Week

- HW3 solution on Piazza soon
- HW4 due Thursday, Oct 31. Start now!
  - See Errata post on Piazza!
- Friday lab **10-11:30 am** in MD G115
- Final Projects - start forming groups!
  - Group size: 3-4 students
  - Proposals due Monday, Nov 11
  - More information coming soon



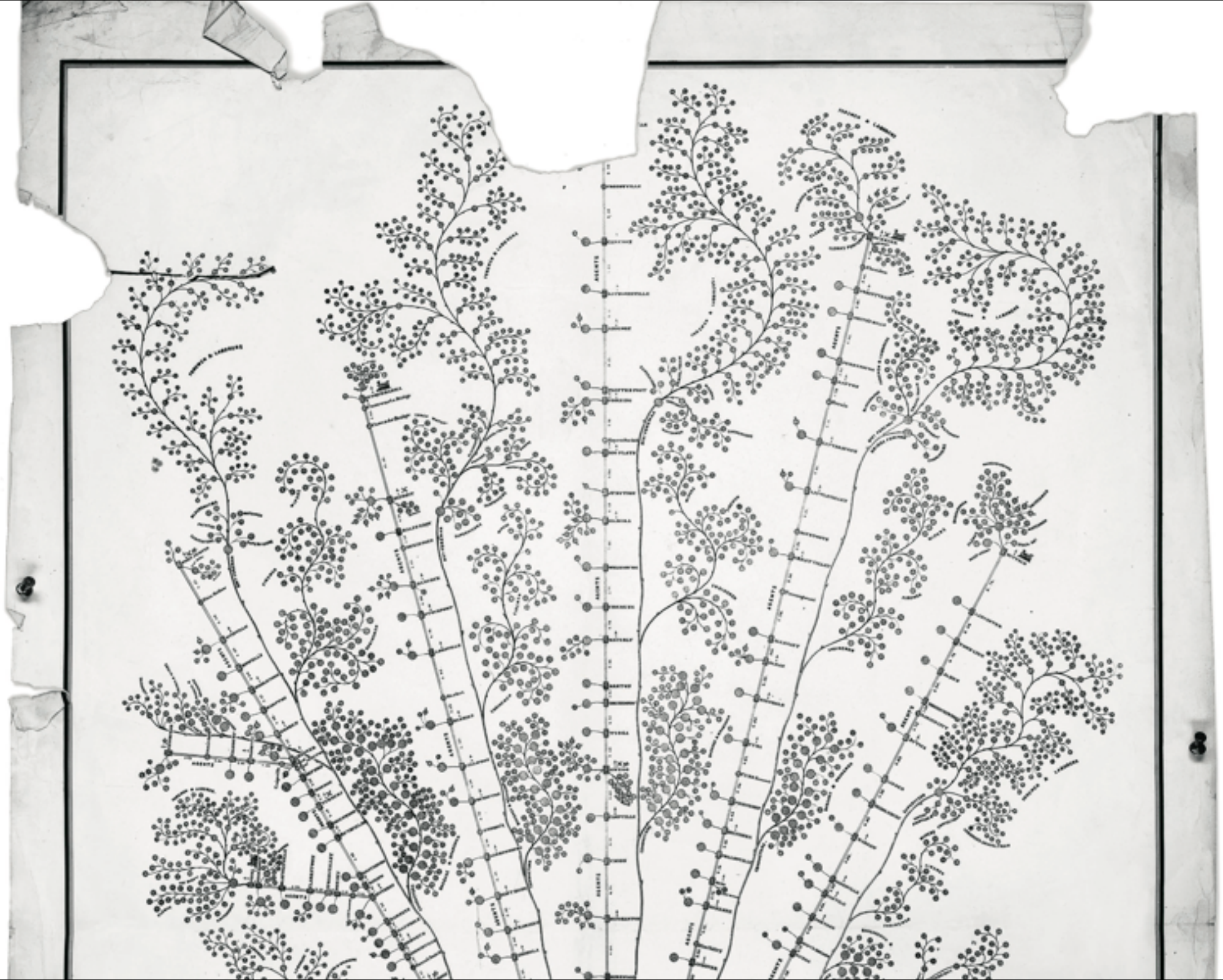


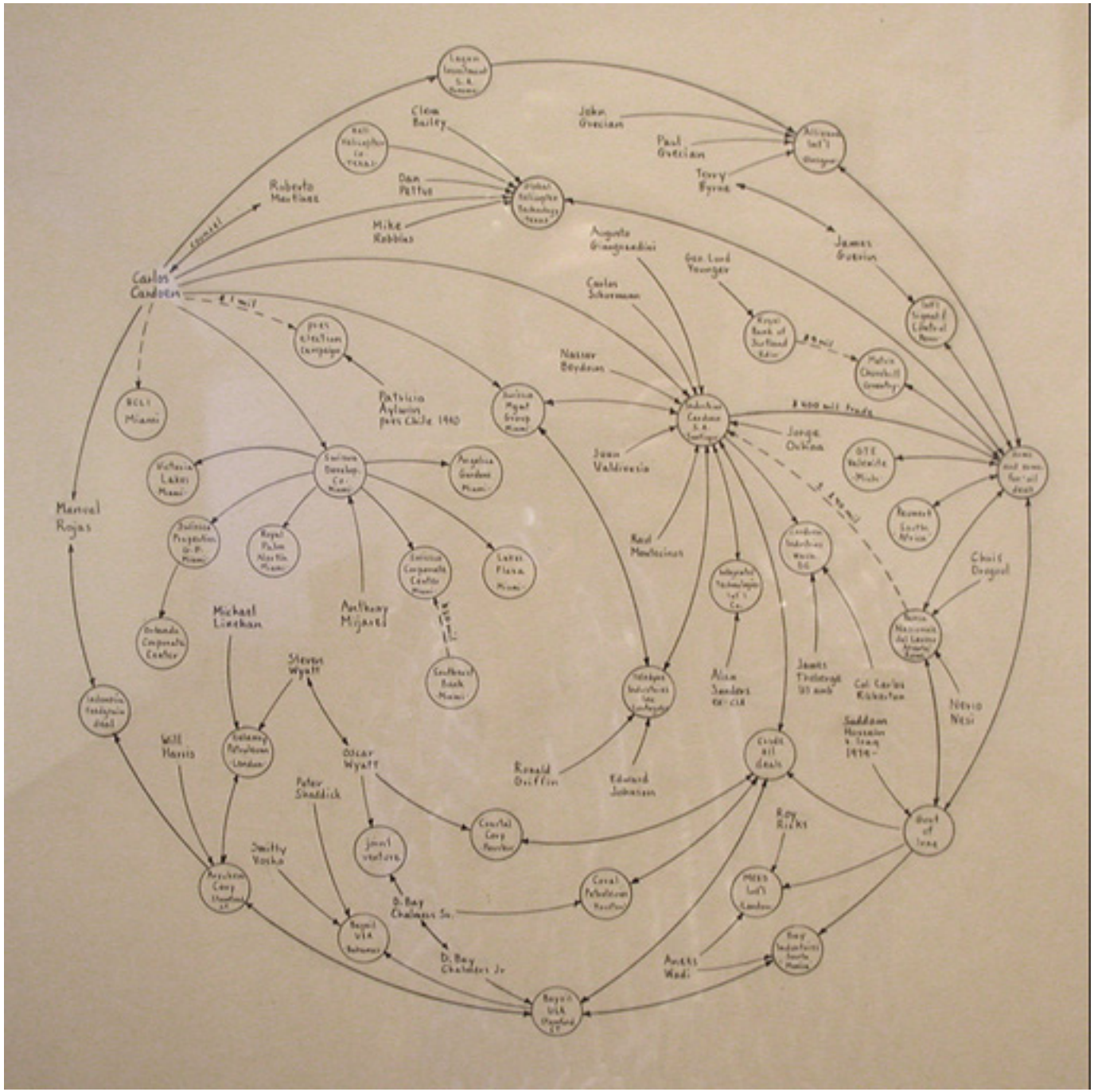
[Godel, Escher, Bach. Hofstadter 1979]



Plan of Organization,
New York and Erie Railroad,
1855

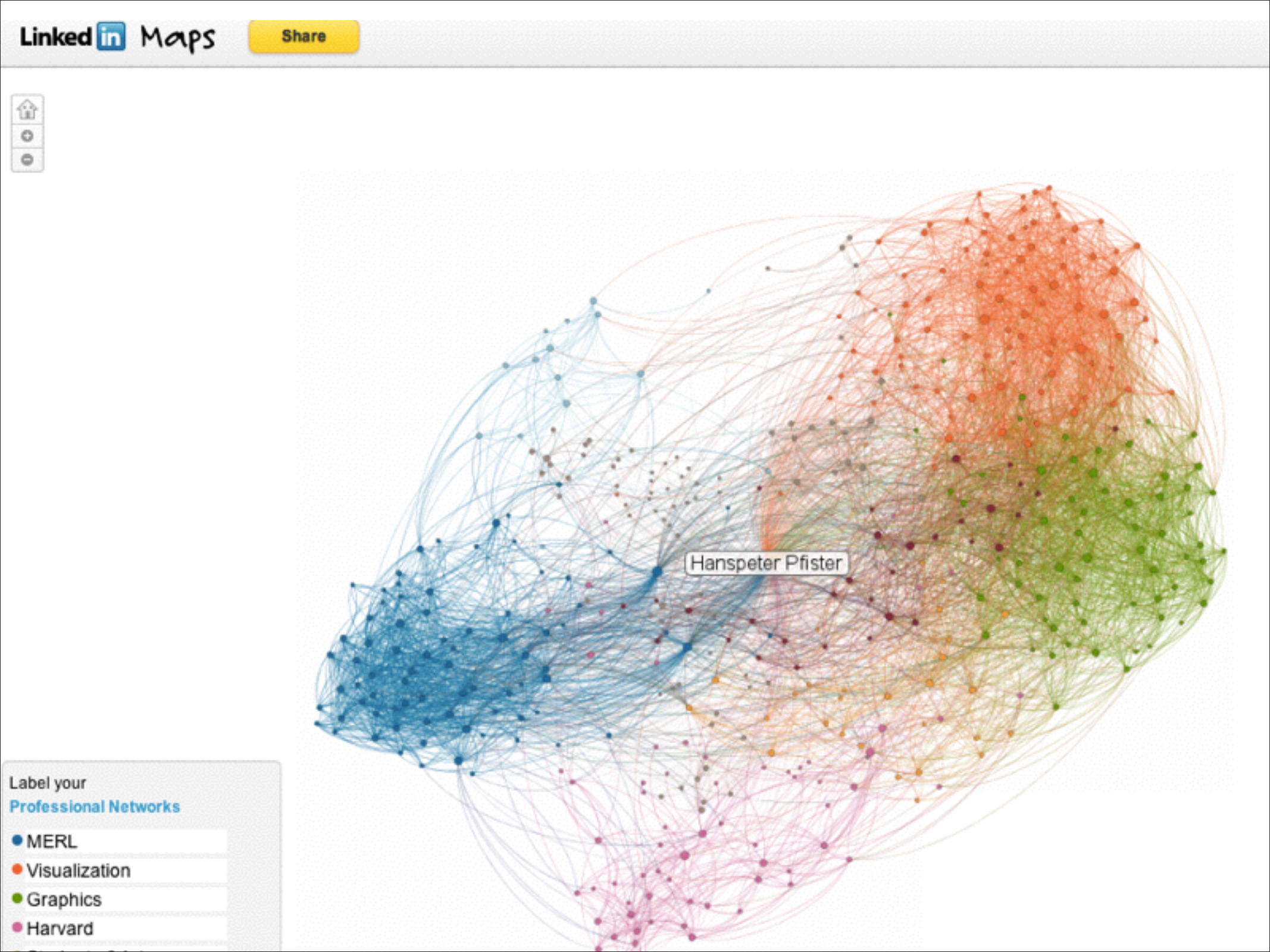
Courtesy of the Geography
and Map Division, Library of Congress.

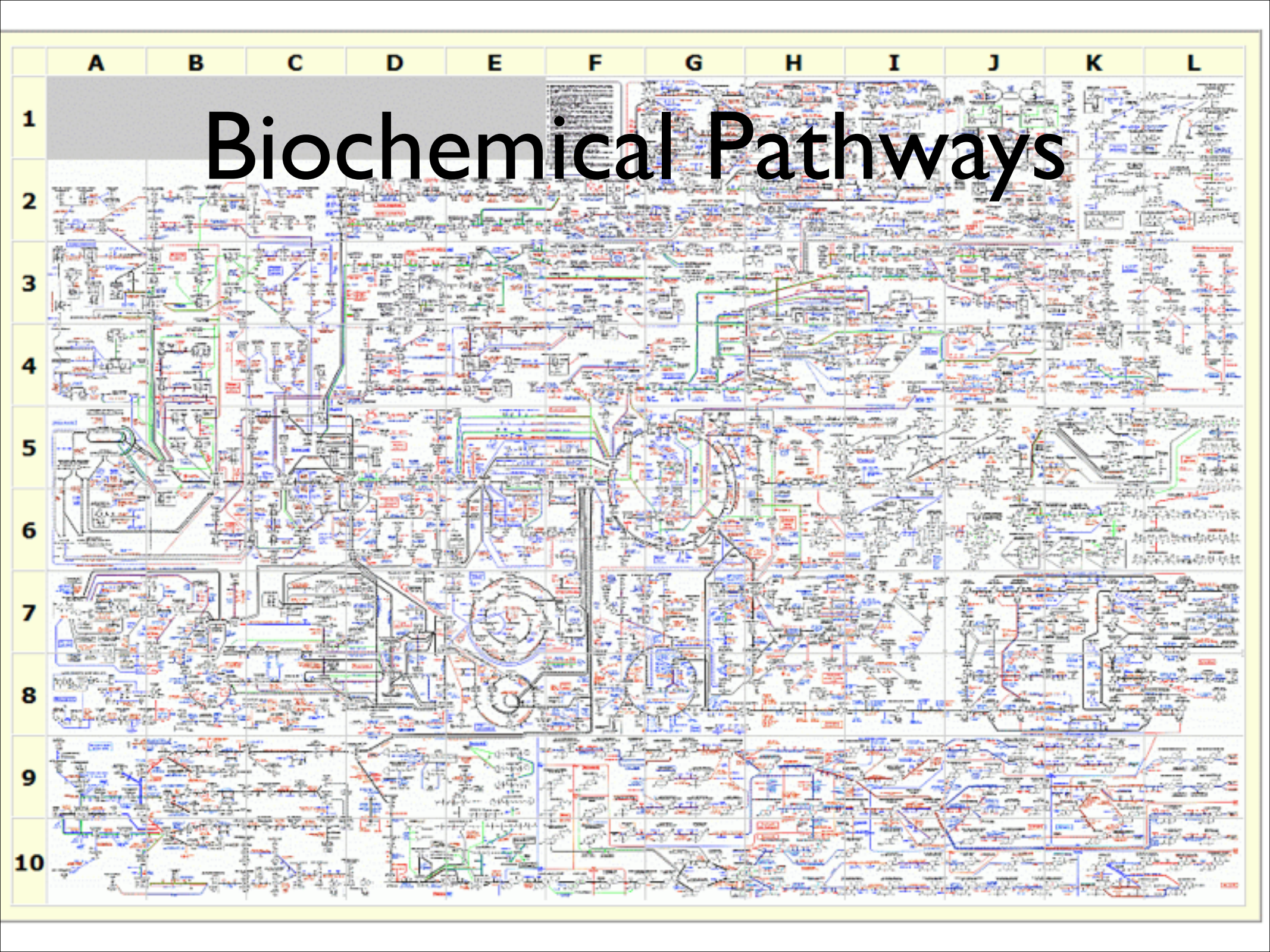




http://benfry.com/exd09/

Mark Lombardi


LinkedIn Maps
Share
Hanspeter Pfister
Label your
Professional Networks
MERL
Visualization
Graphics
Harvard

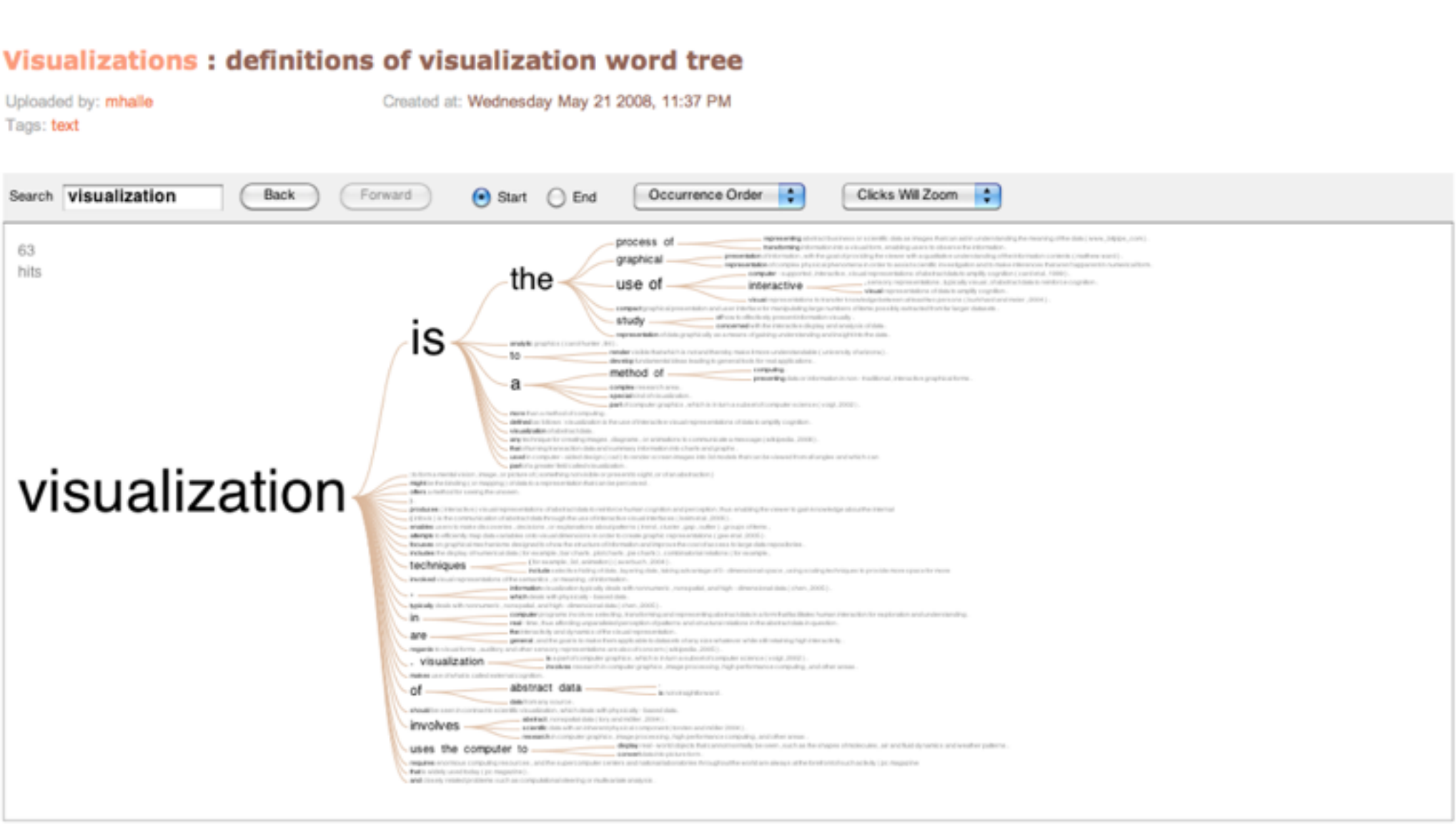

Biochemical Pathways


facebook


December 2010

Trees

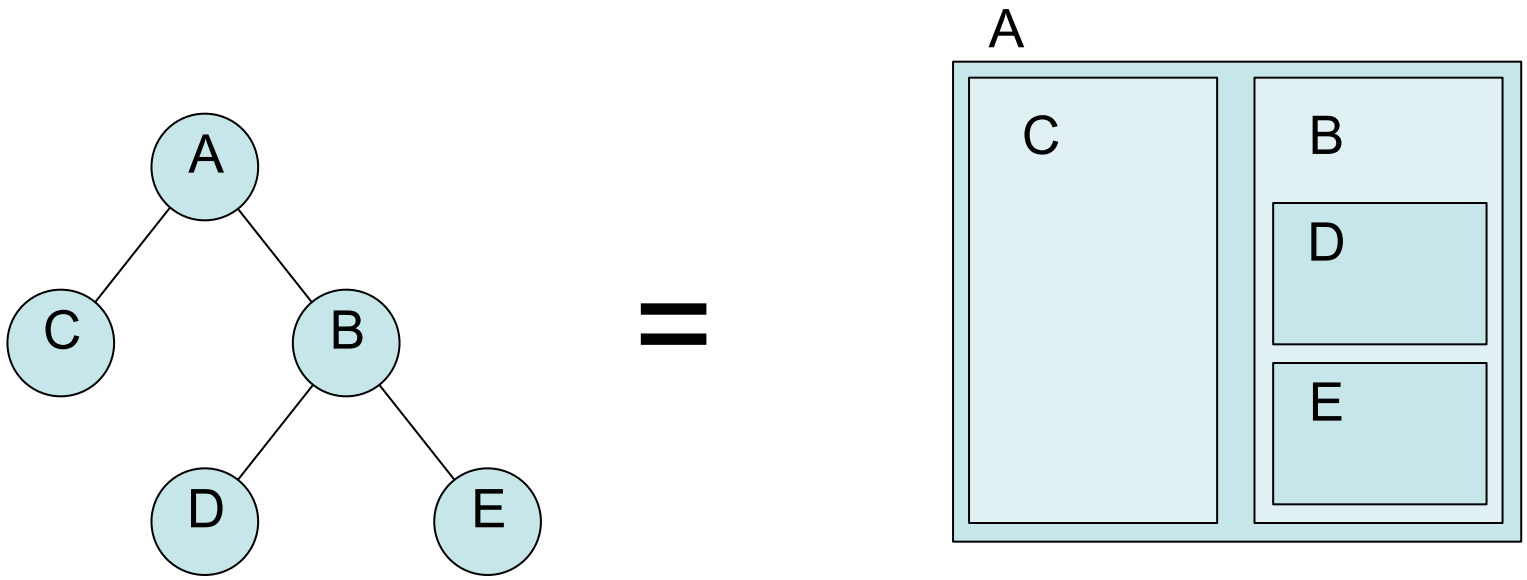
Indented Trees



Node-Link Trees

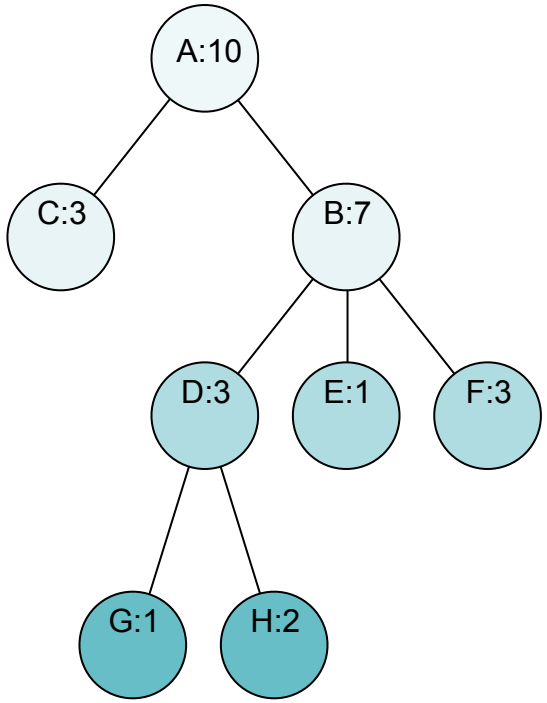
Enclosure

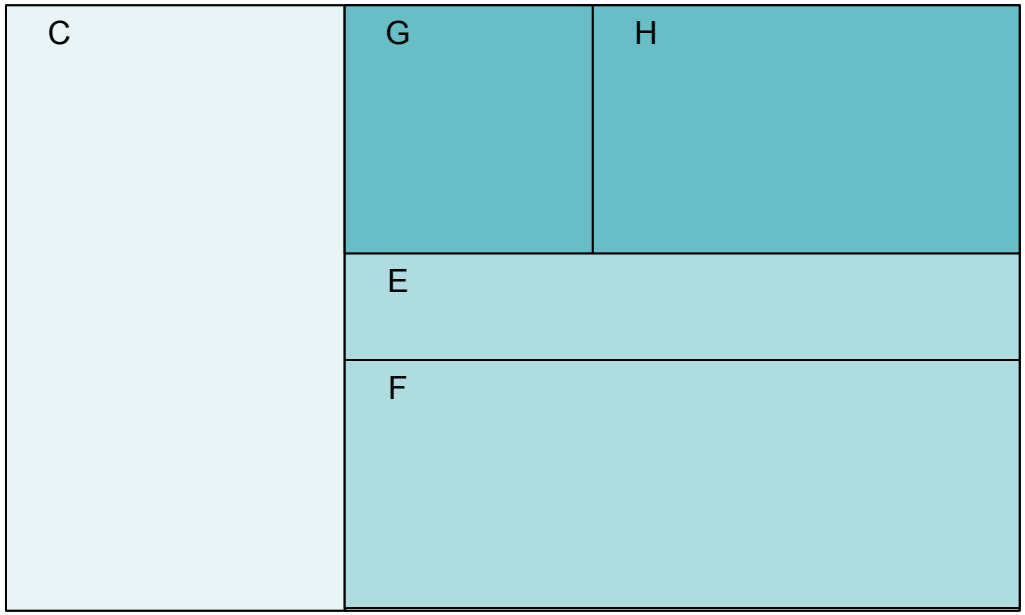
Indicate parent – child relationship by visually enclosing children within parent



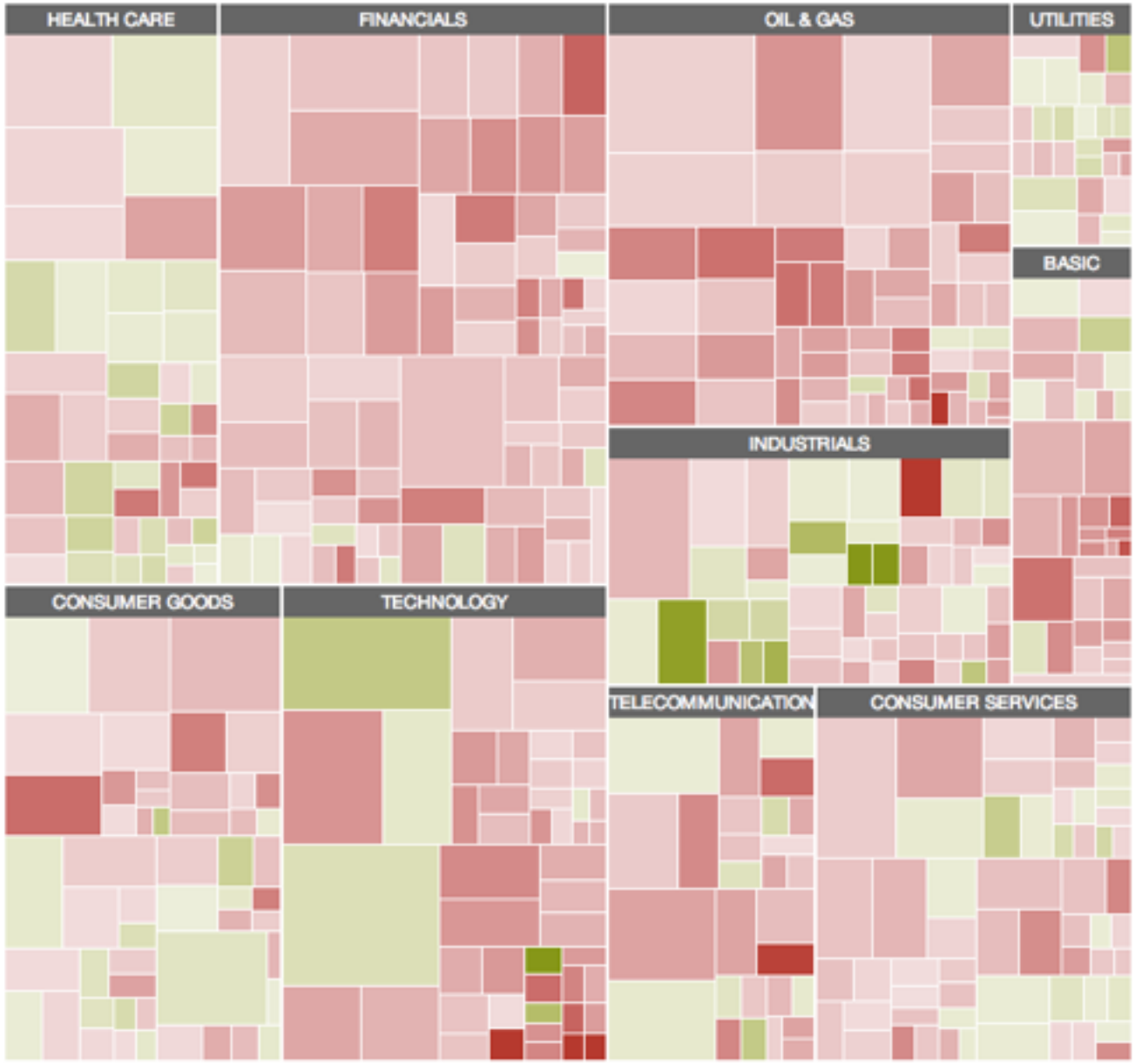
Treemaps

- Assume each leaf node has an associated size (i.e. files on disk, or salaries in a orgchart)
- Size of parent node is the sum of its children.





Map of the Market



DJIA	-0.35%	Nasdaq -0.57%	S&P 500 -0.47%
	54.33 ↓	22.50 ↓	8.29 ↓
	15413.33	3907.07	1746.38

MarketWatch News Viewer

5:18p **AT&T earnings rise in third quarter** MARKETWATCH
T

5:08p **Visa boosts dividend 21%** MARKETWATCH
V

5:07p **Citrix profit falls 1.9%, costs mask sales growth** MARKETWATCH
CTXS

4:59p **Symantec profit up 28%, outlook cut** MARKETWATCH
SYMC

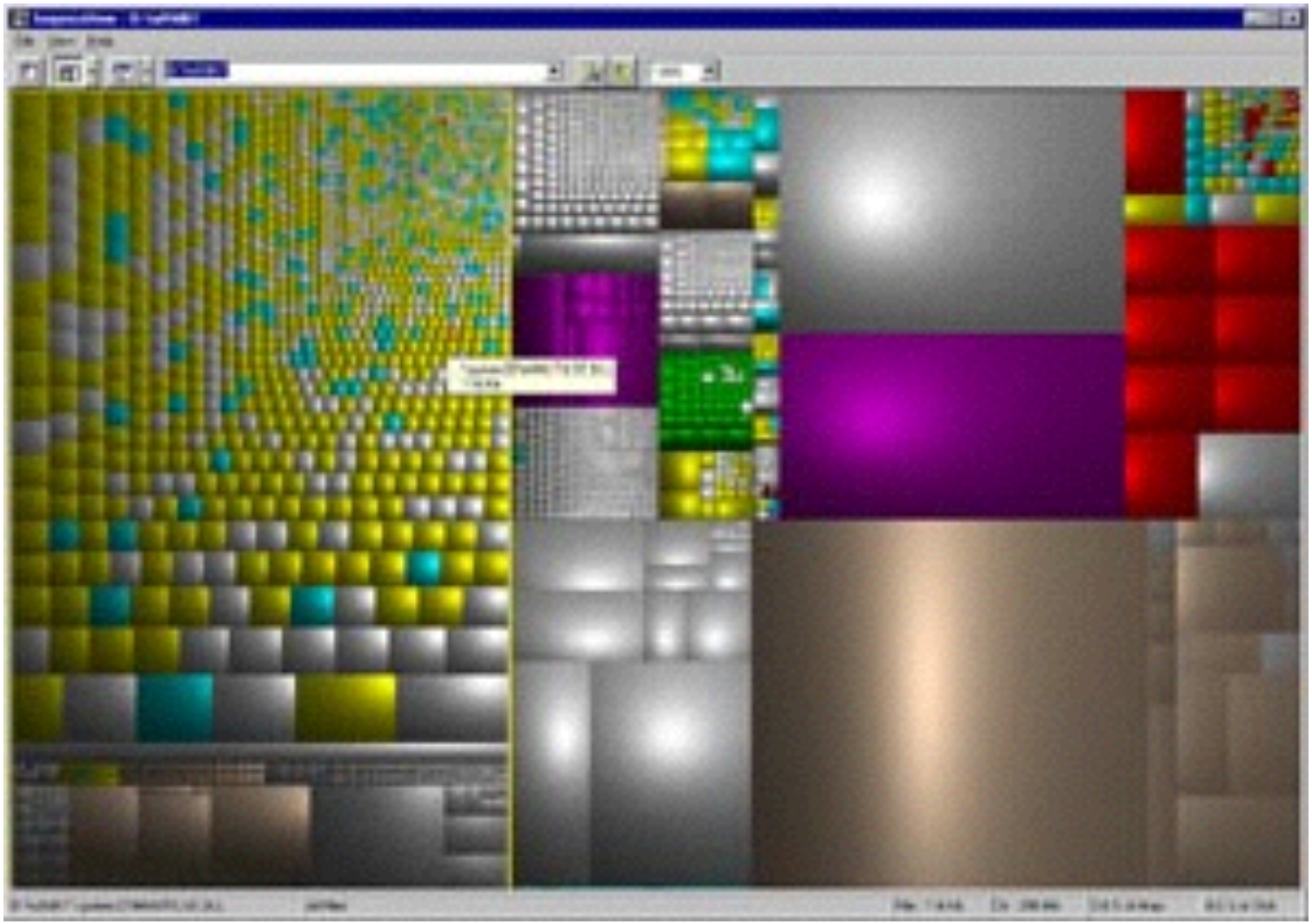
4:53p **Wal-Mart de Mexico posts meager financial growth** MARKETWATCH
WMT

4:47p **E*Trade swings to profit; sells market-maker unit** MARKETWATCH
SCHW

4:27p **BofA liable for Countrywide fraud, jury finds** MARKETWATCH
BAC

3:41p **AT&T, techs report earnings after bell** MARKETWATCH

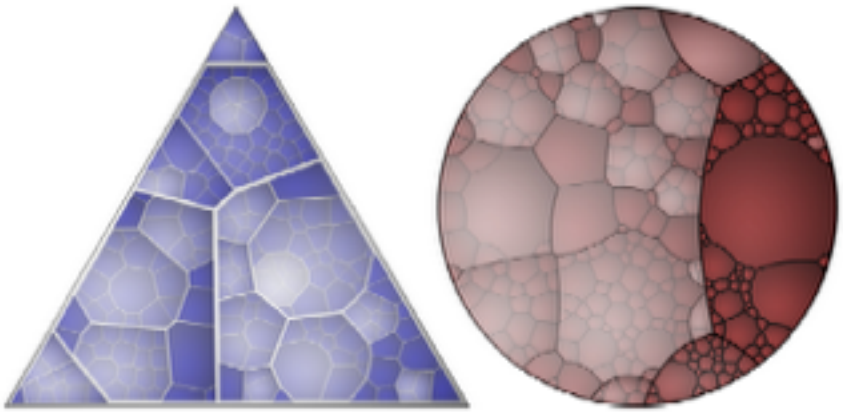
Sequoia View



http://w3.win.tue.nl/nl/onderzoek/onderzoek_informatica/visualization/sequoiaview/

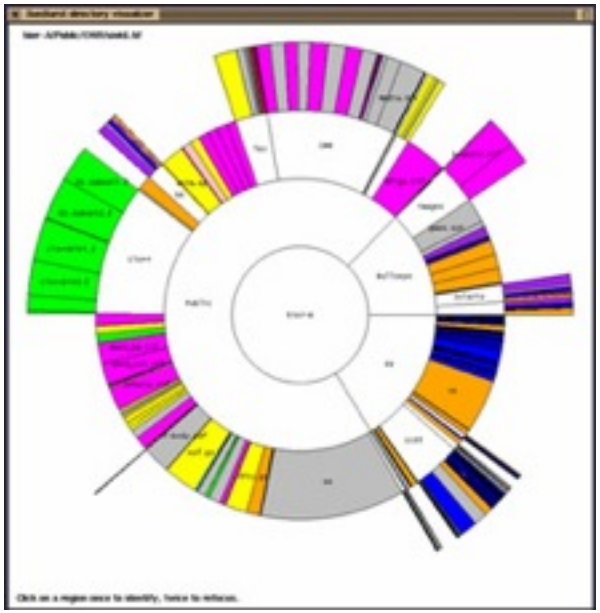
Treemap Problems

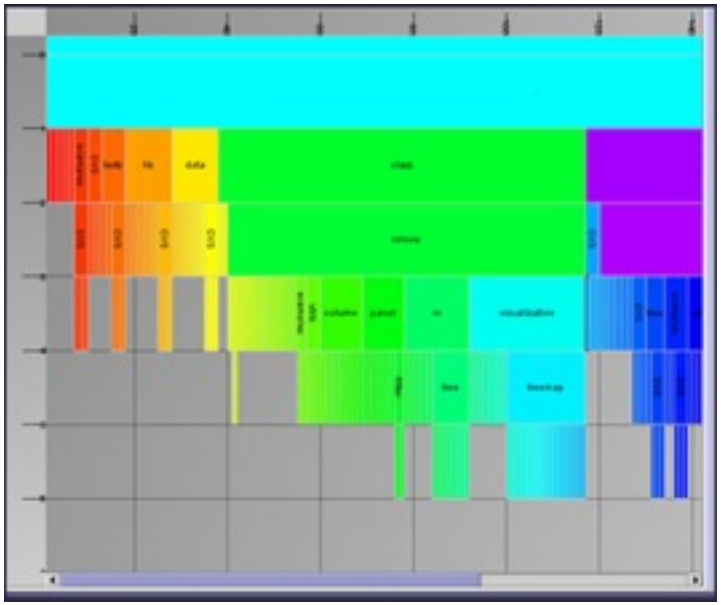
- Recursive slice-and-dice subdivision pattern leads to long and thin rectangles.
- Impossible to interact with internal nodes



Layering

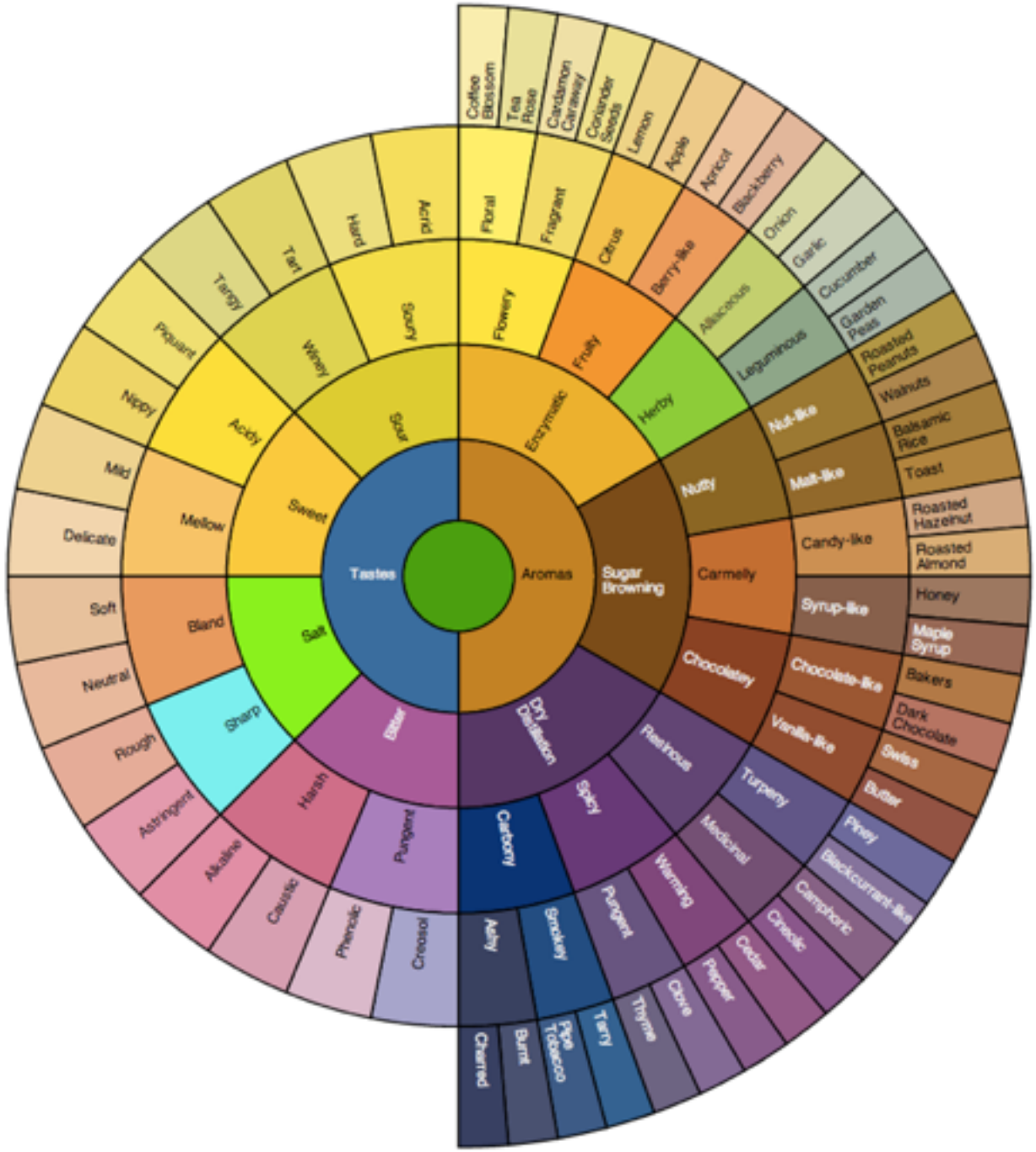
- Similar to node link layouts without edges
- Depth on one axis, recursive layout on the other





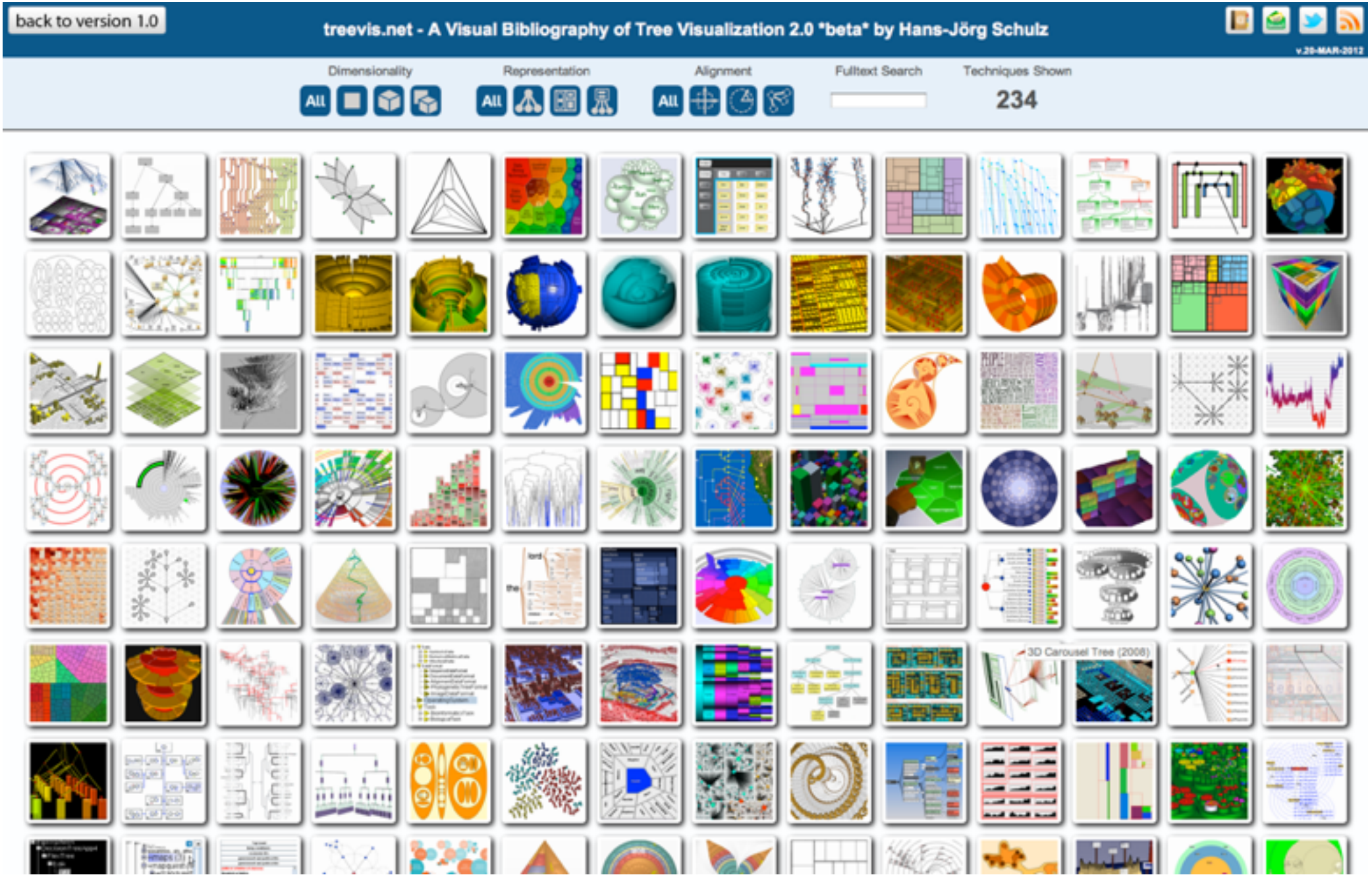
2008												
Q1			Q2			Q3			Q4			
Jan	Feb	March	April	May	June	July	August	September	October	November	December	
7856	987	89	456	234	4567	5621	542	451	14	125	62	
48	6542	55	654	54	6	64	65	62	245	654	24	

Coffee Flavour Wheel

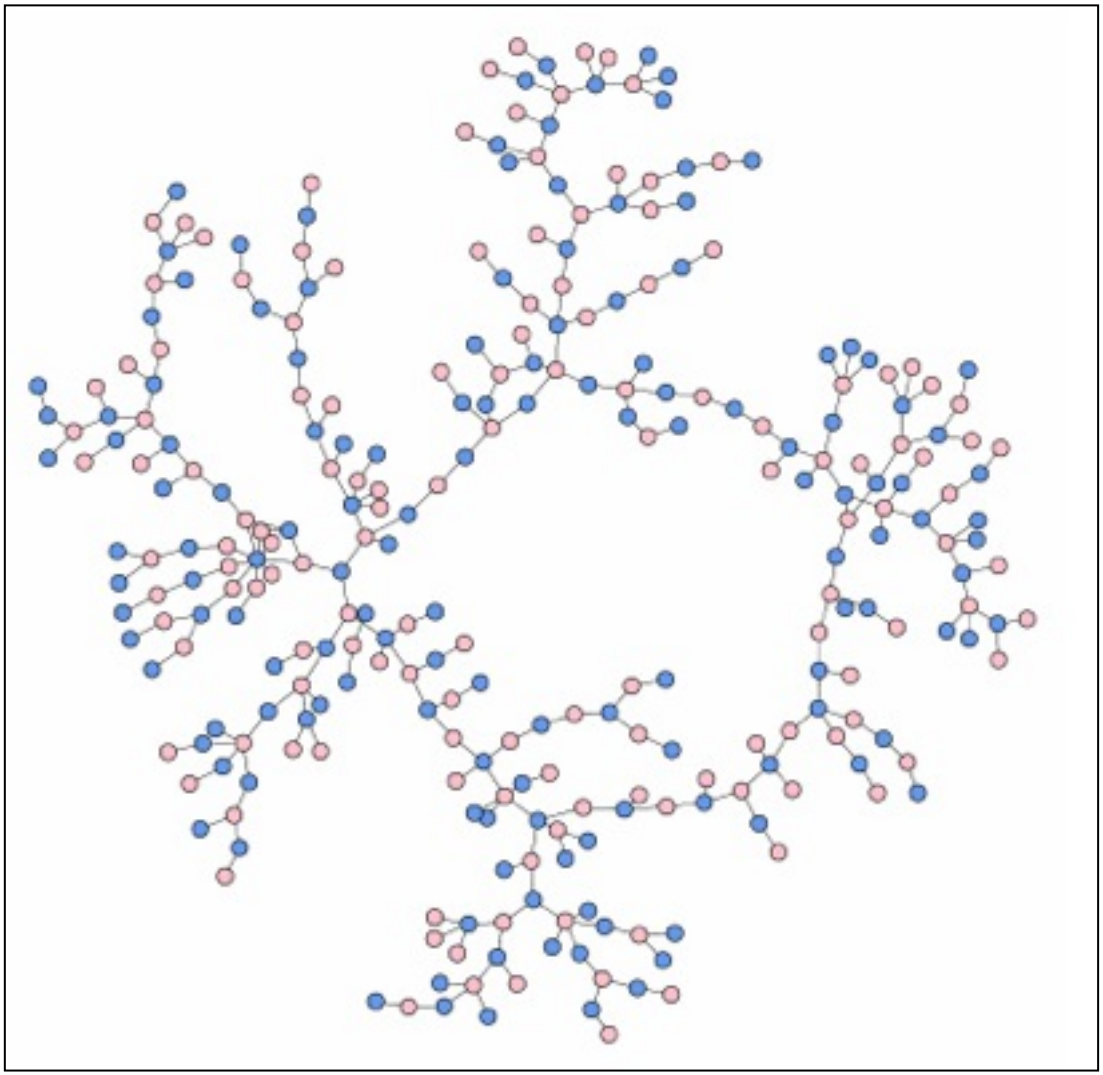


Click to zoom!

d3


back to version 1.0
treevis.net - A Visual Bibliography of Tree Visualization 2.0 *beta* by Hans-Jörg Schulz
v.20-MAR-2012
Dimensionality
All
Representation
All
Alignment
All
Fulltext Search
Techniques Shown
234
3D Carousel Tree (2008)

Networks

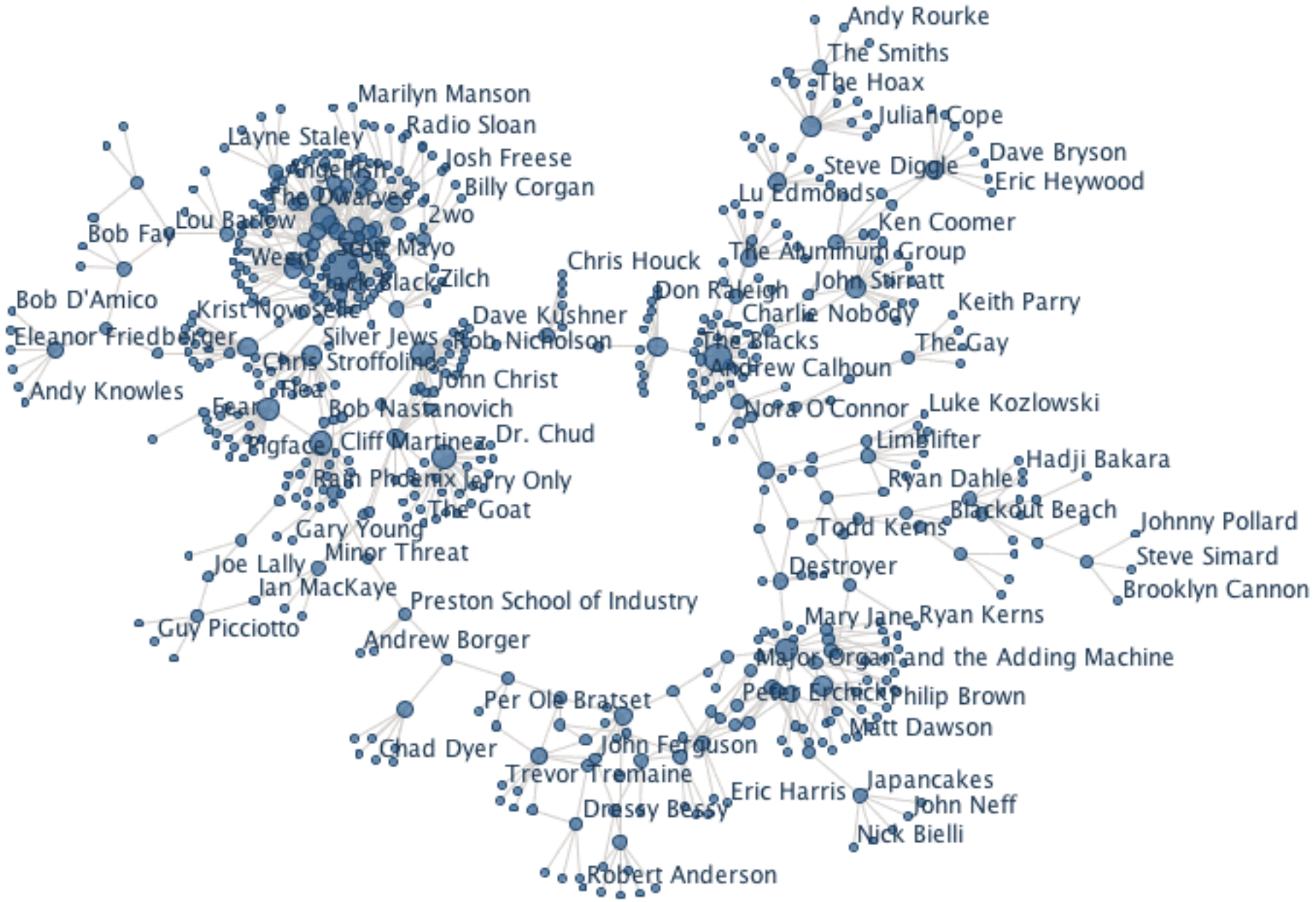


High school dating network

Visualizations : Music Networks, as of 21 March 2007

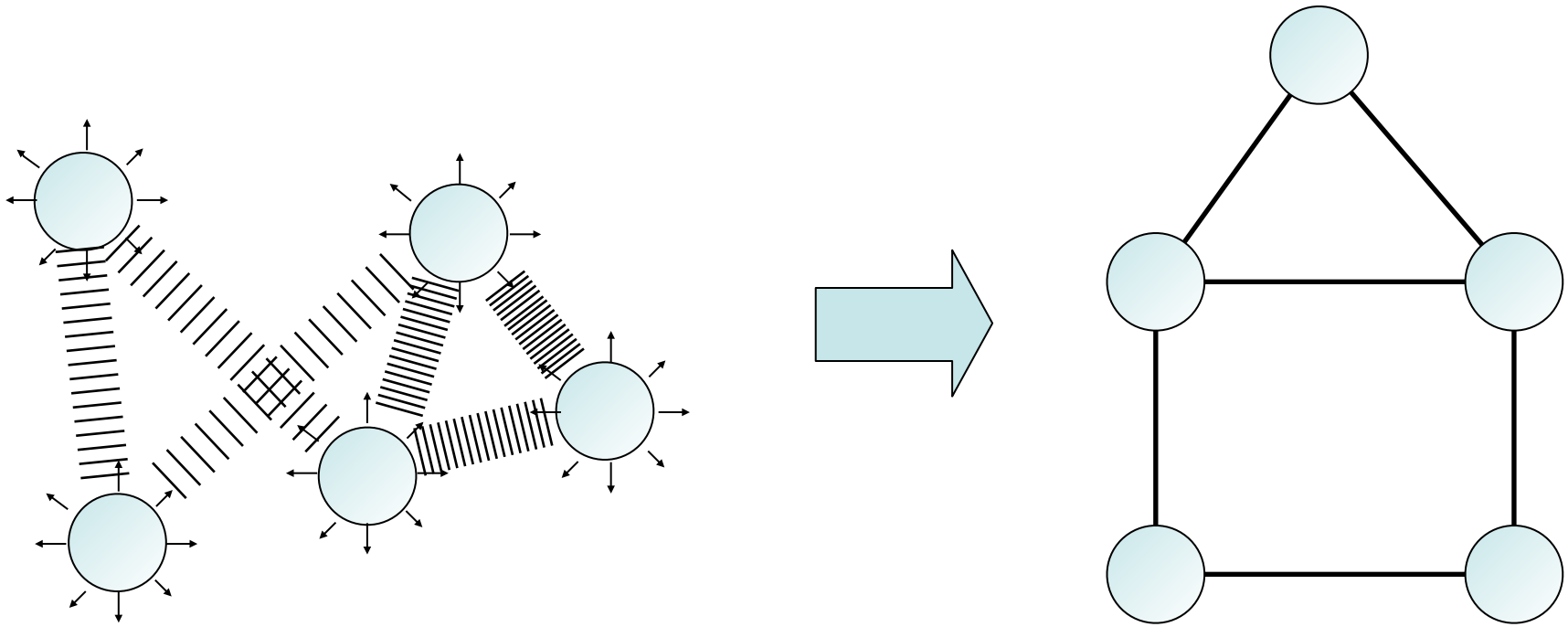
Uploaded by: brainwidth

Created at: Wednesday March 21 2007, 11:16 AM



Force Directed Layouts

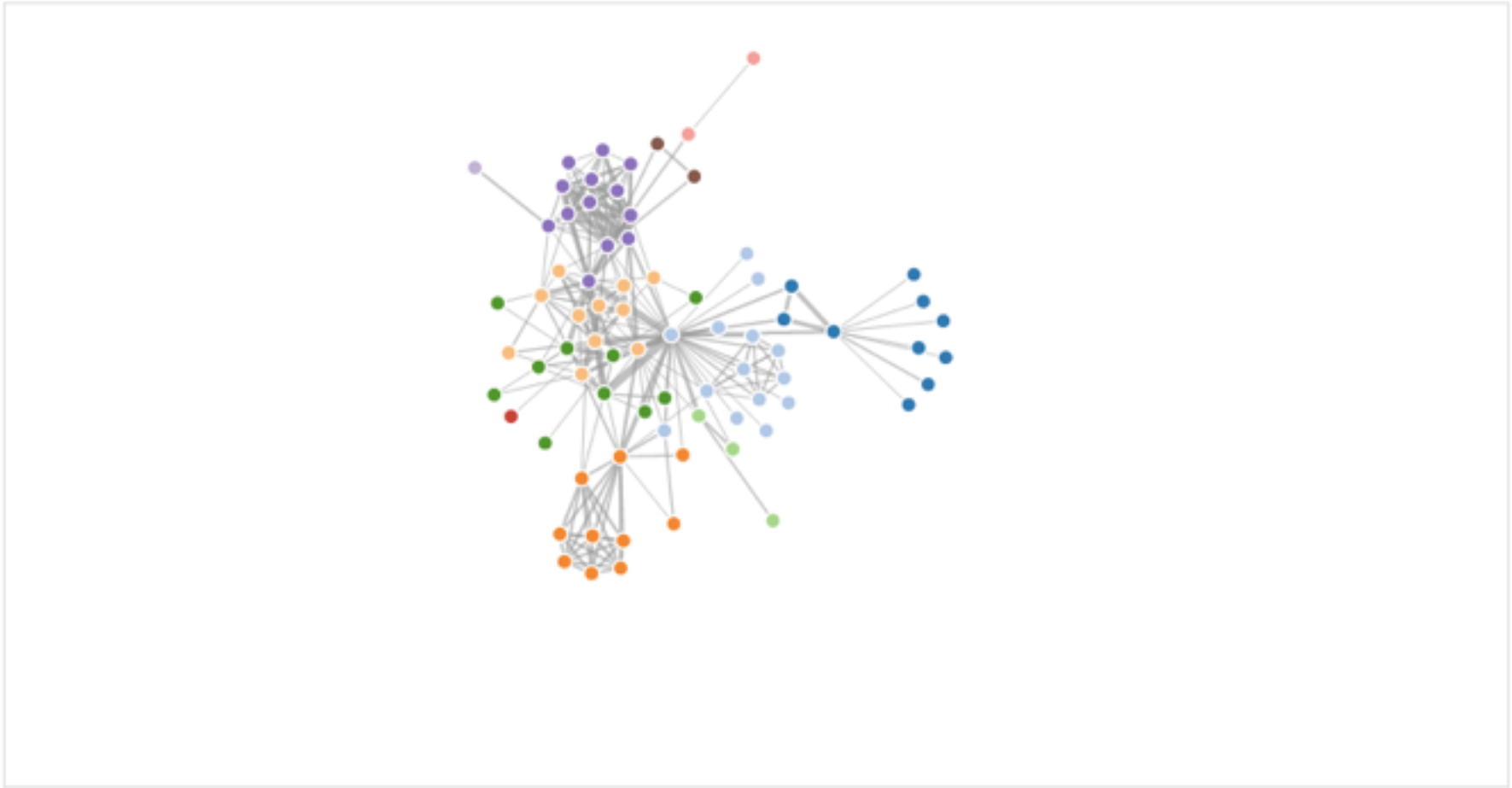
Physics model, edges = springs, nodes = repulsive magnets



mbostock's block #4062045

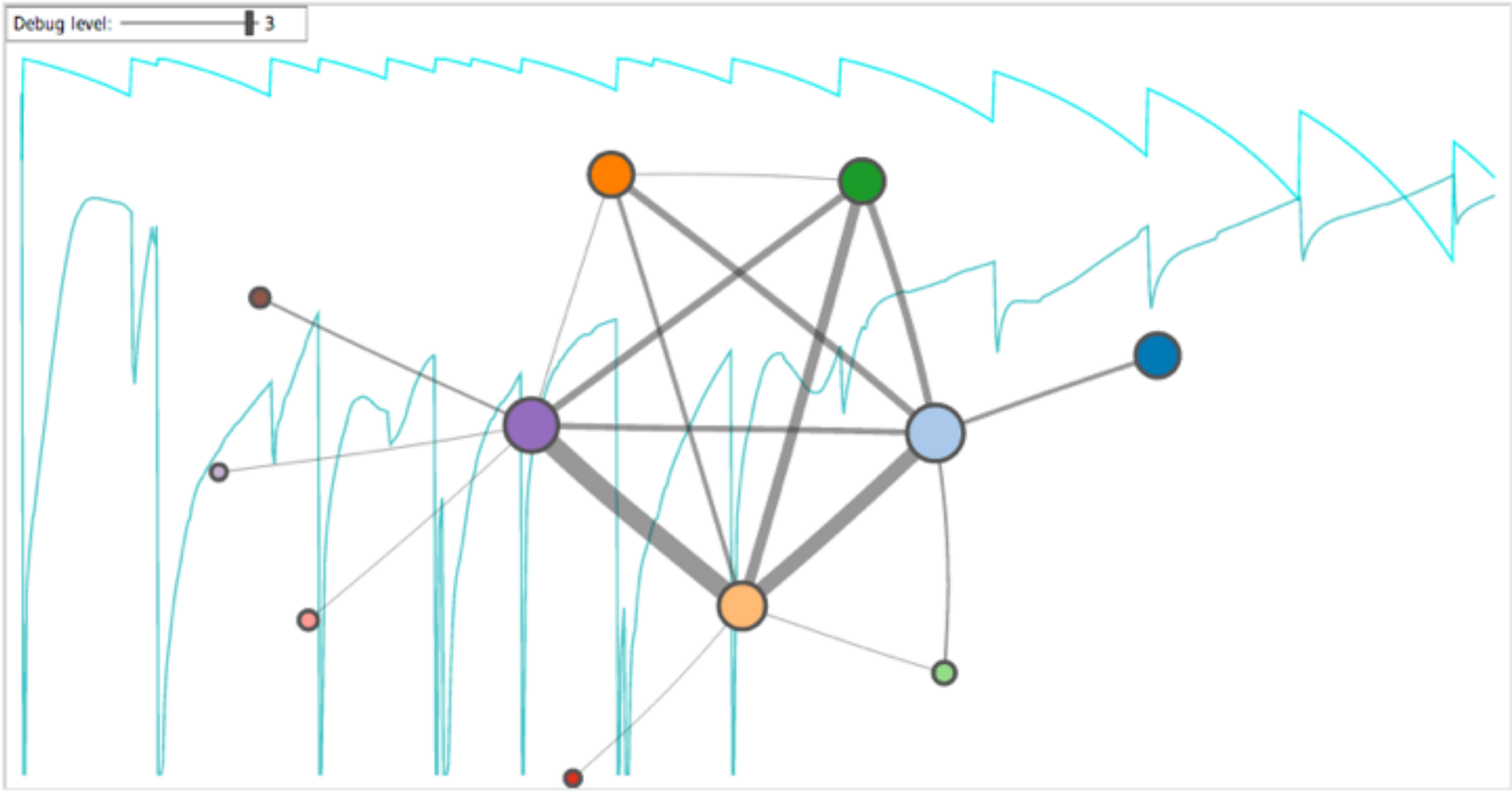
Force-Directed Graph

March 21, 2013



d3.js: force layout with self-referencing links

March 1, 2013



mentionmapp

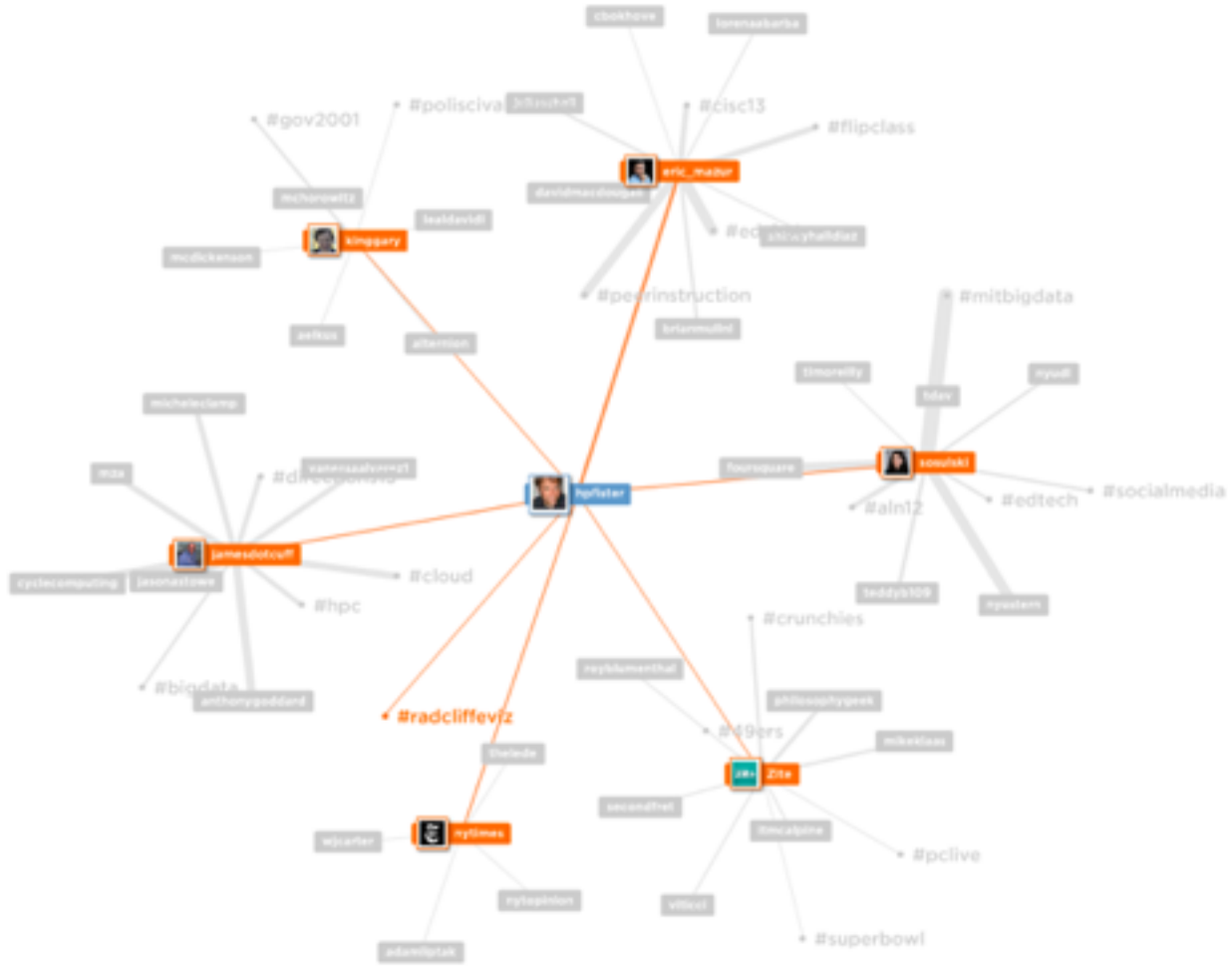
mentionmapp : hpfister

hpfister

You're looking at a map of **mentions**. Each user is connected to the people and hashtags they mentioned the most in recent tweets. Click a node to explore its neighborhood.

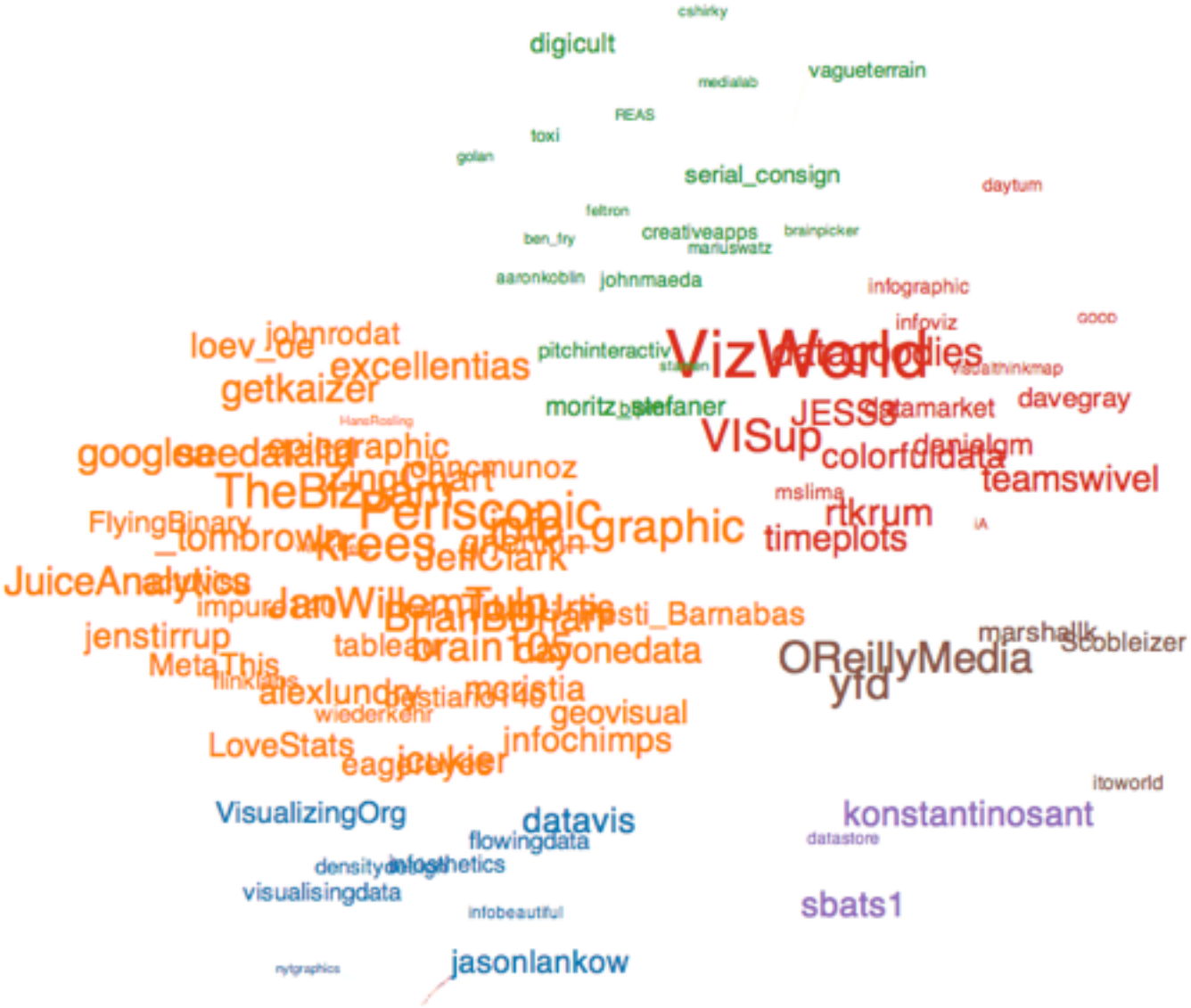
id	hpfister
name	hpfister
location	Cambridge, MA, USA
profile	http://twitter.com/hpfister
url	

Feedback Retweet About



Union of Top Nodes by Eigenvector Centrality, Degree, Betweenness

Nodes shown represent a high-scoring subset of the full 1644 node dataset from Moritz Stefaner's crawl of infovis tweeters in Summer 2011.
Nodes are colored by partition. Click on a node to see stats and follower relations.



Data Subset Stats

Nodes: 95 Edges: 3636
N Used: 50

Data shown is result of a union of the top N nodes having highest eigenvector centrality, degree, and betweenness, computed in the javascript off the full dataset.

Size Node Text By:

Eigenvector Centrality

Selected Node:

Click on a node!

Built with D3 by Lynn Cherny from NetworkX analysis with accompanying talk slides and blog post.

vizster

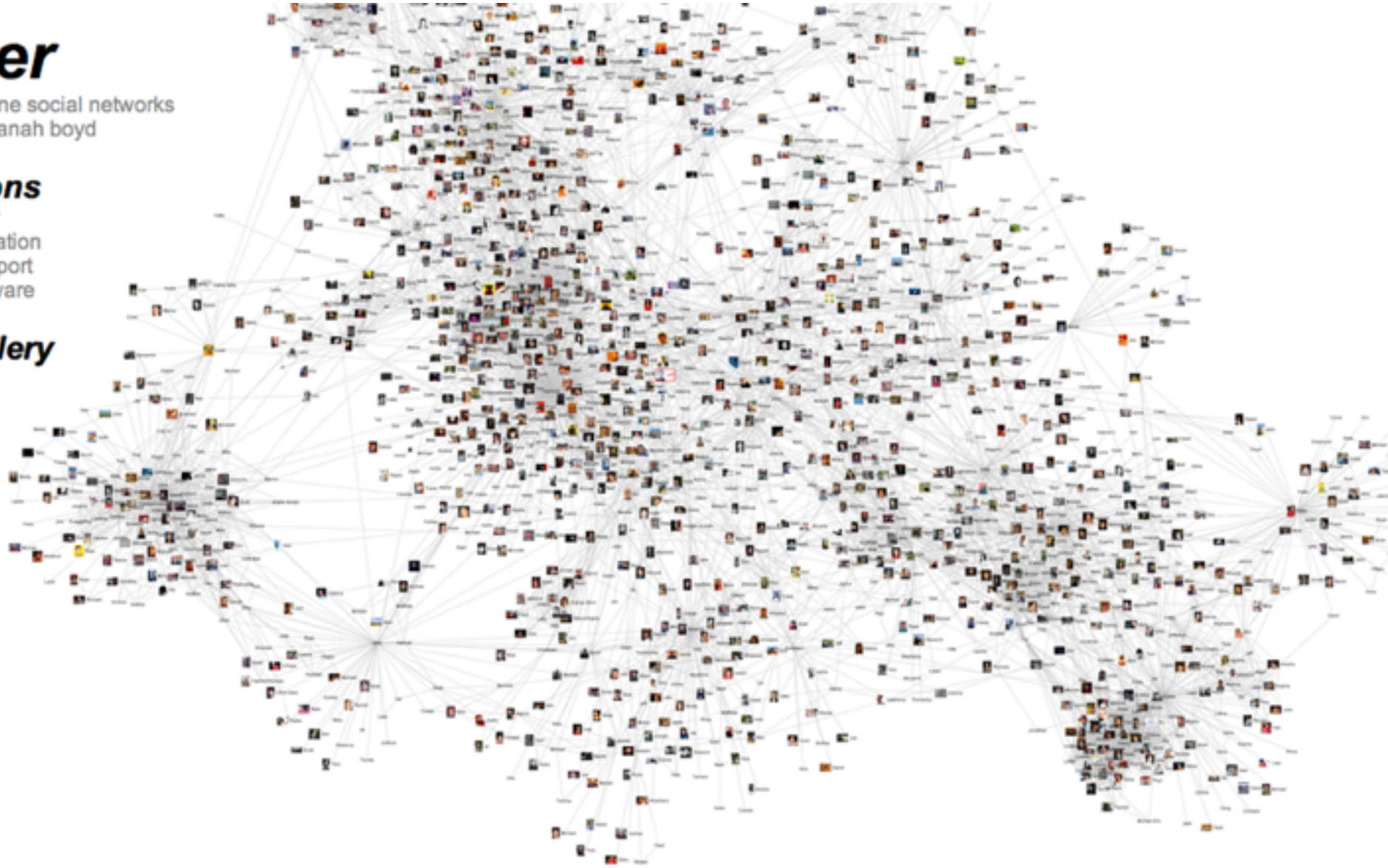
visualizing online social networks
jeffrey heer + danah boyd

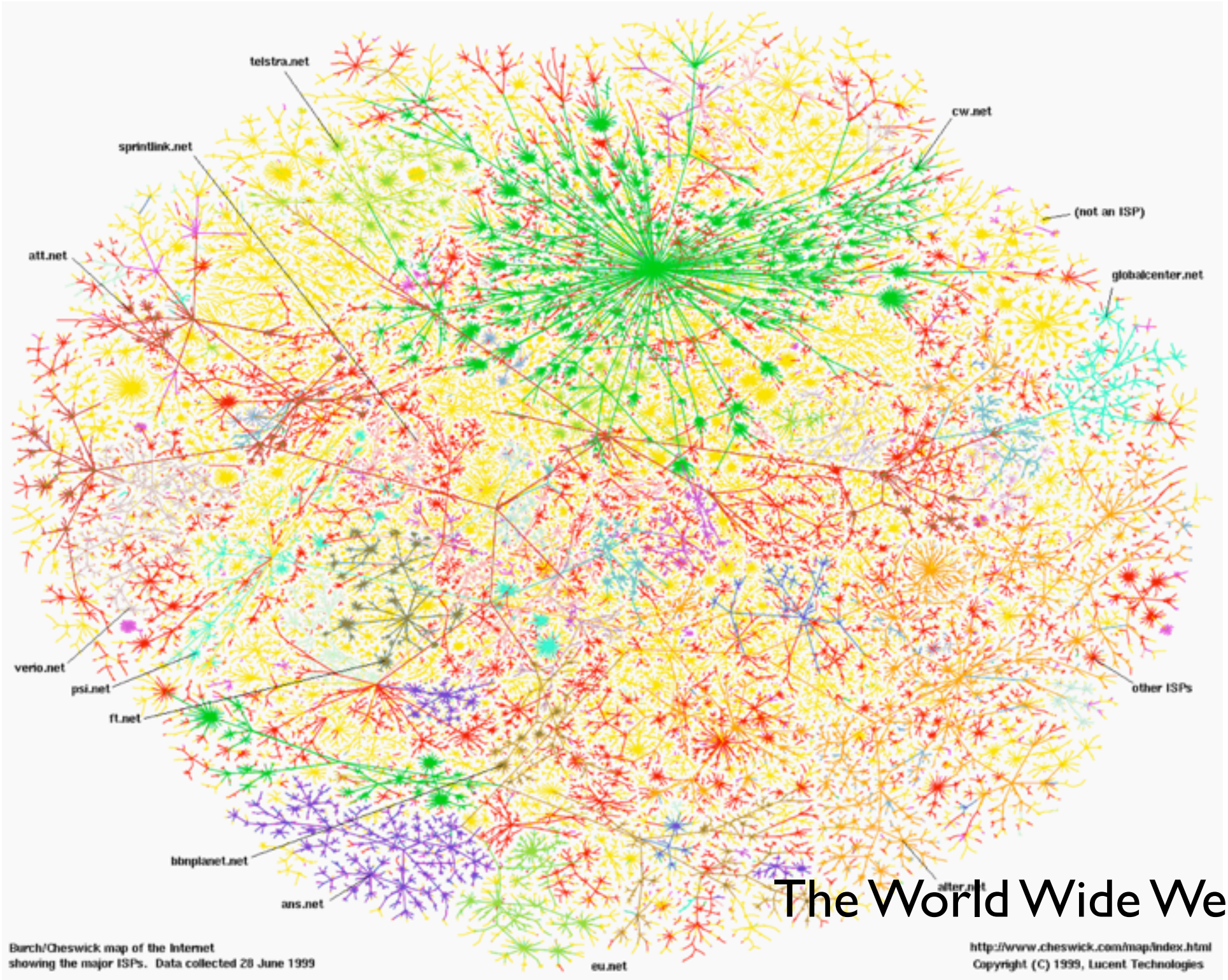
publications

research paper
video demonstation
early design report
download software

photo gallery

egocentric
community
linkage
search
x-ray 1
x-ray 2

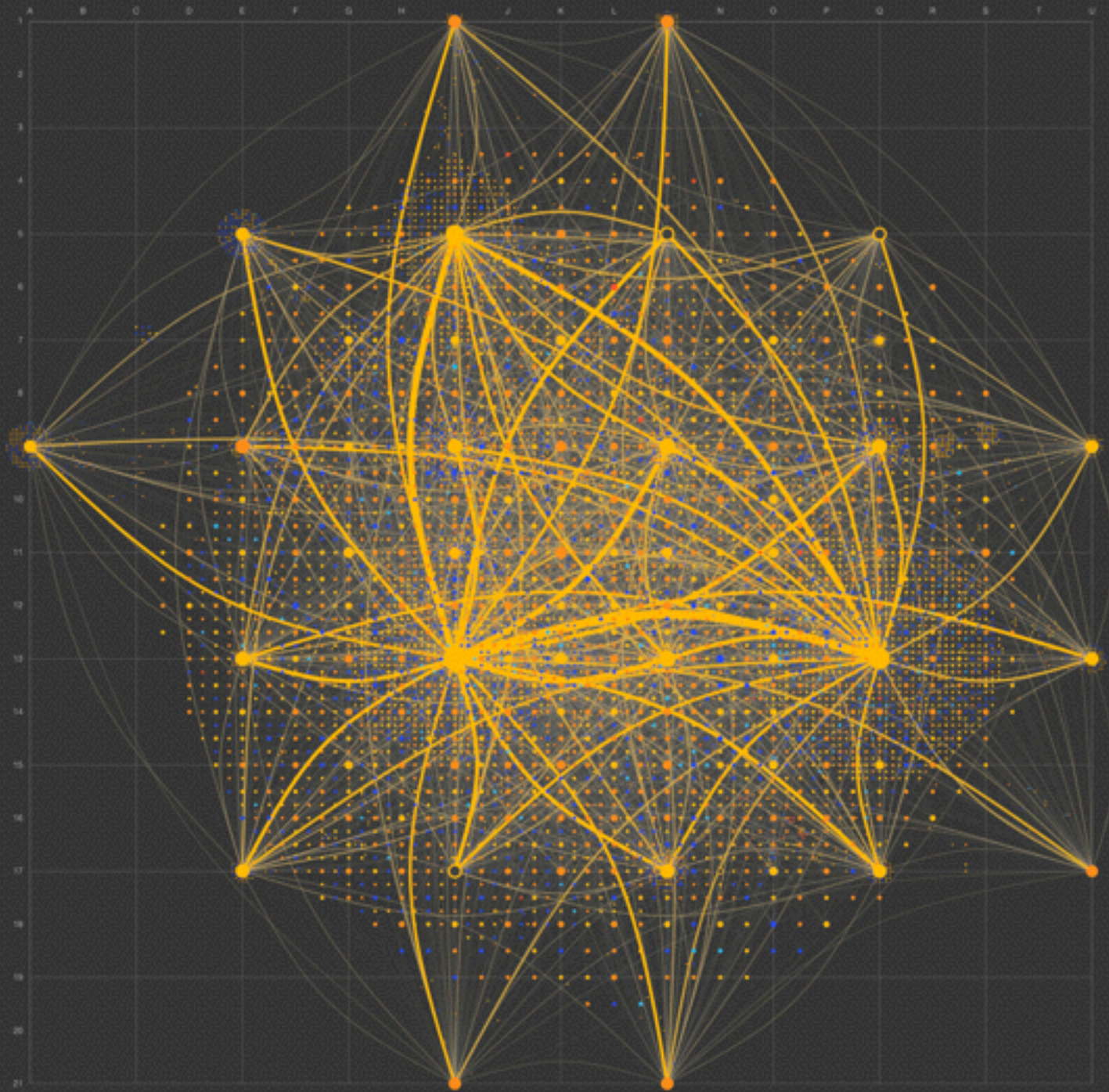




The World Wide Web

The Internet

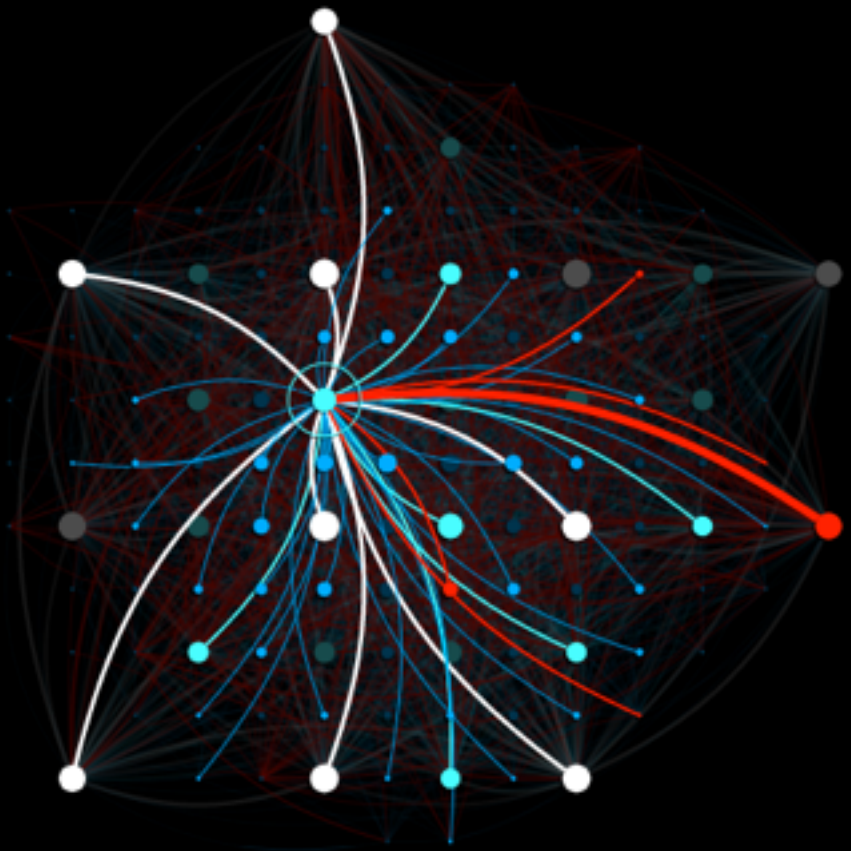
Topology of Autonomous Systems, 2011.01.02



Twitter Connections Before and After Eyeo 2012

Follow Graph for Conference Attendees Registered at lanyrd.com

Visualization by Jeff Johnston



Followers of

factoryfactory

seattle, usa

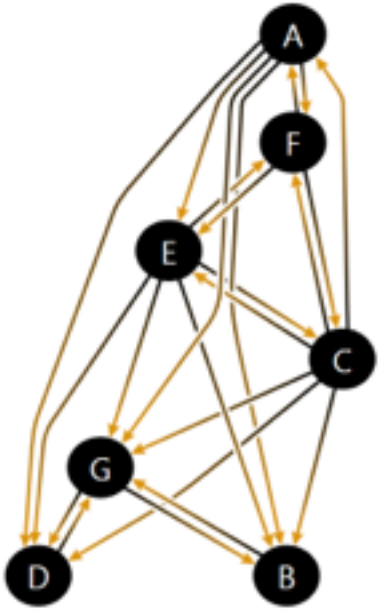
Followers: 1269
Followers Among Attendees: 39%
Following: 244
Following Among Attendees: 22%

Mouse over to explore connections
Click to toggle between followers and following

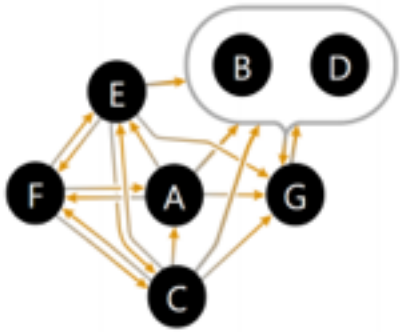
- Most Central
- Moderately Central
- Least Central
- New Connections

Edge Compression Techniques for Visualization of Dense Directed Graphs

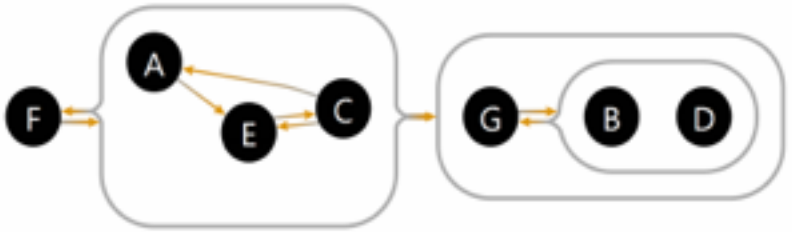
Tim Dwyer, Nathalie Henry Riche, Kim Marriott, Christopher Mears



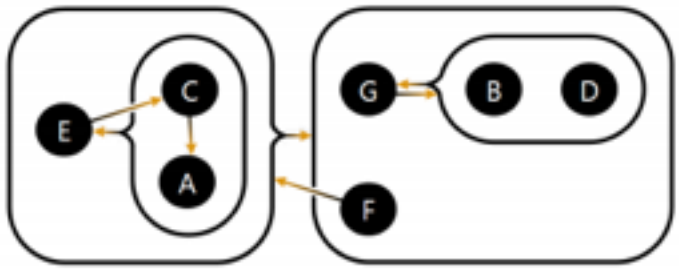
(a) A standard "flat" node-link diagram of a graph with 23 directed edges.



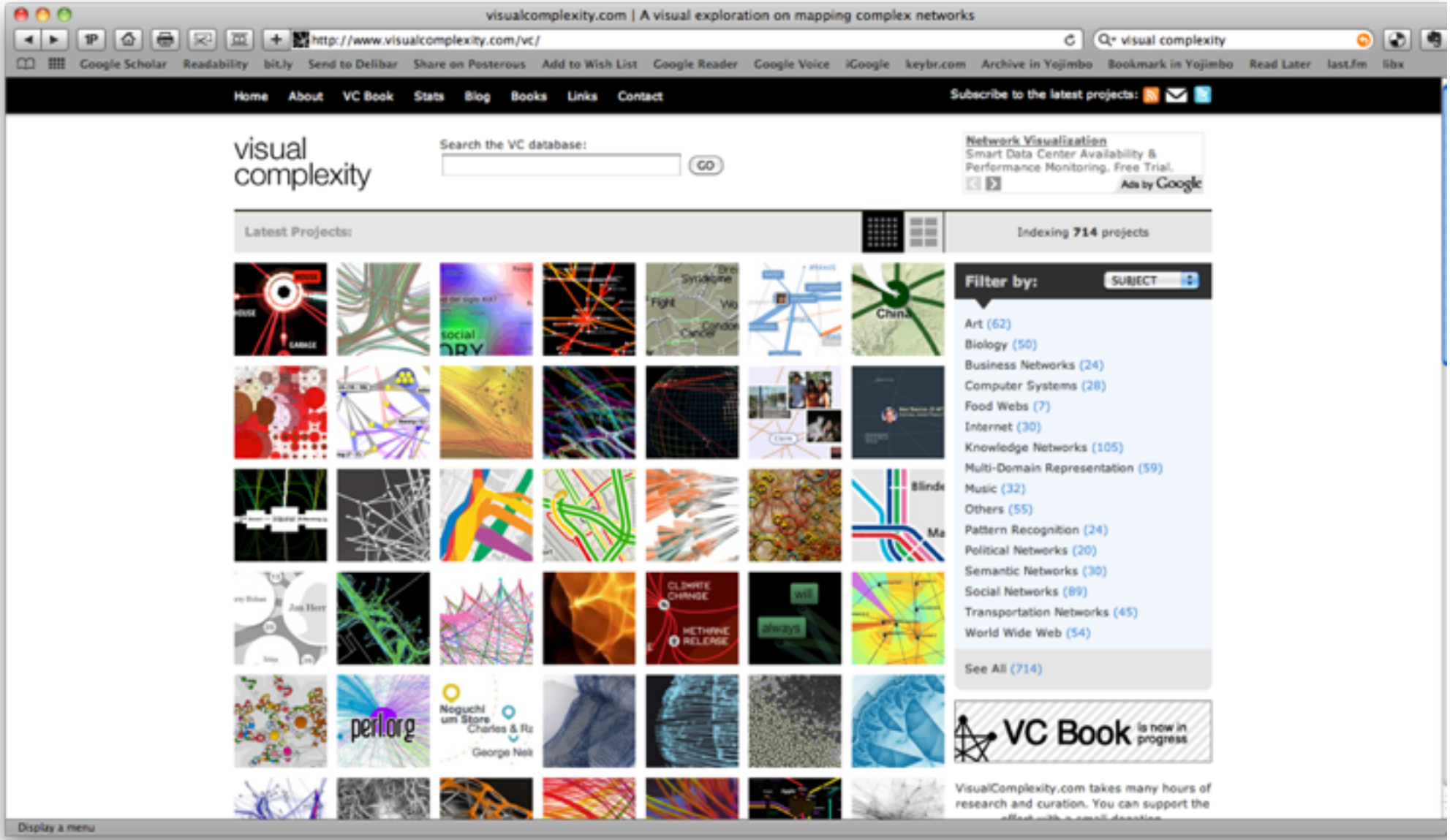
(b) Since B and D have exactly the same sets of neighbors they can be grouped leaving 18 edges using a simple *Neighbor Matching*.



(c) A *Modular Decomposition* allows internal structure within modules and nesting. Nine edges remain.



(d) A *Power-Graph Decomposition* further relaxes the definition of a module to allow edges to cross module boundaries, allowing the same graph to be drawn with only 7 edges.


visualcomplexity.com | A visual exploration on mapping complex networks
http://www.visualcomplexity.com/vc/
visual complexity
Google Scholar Readability bit.ly Send to Delibar Share on Posterous Add to Wish List Google Reader Google Voice iGoogle keybr.com Archive in Yojimbo Bookmark in Yojimbo Read Later last.fm libx
Home About VC Book Stats Blog Books Links Contact
Subscribe to the latest projects:
visual complexity
Search the VC database:
GO
Network Visualization
Smart Data Center Availability & Performance Monitoring. Free Trial.
Ads by Google
Latest Projects:
Indexing 714 projects
Filter by:
SUBJECT
Art (62)
Biology (50)
Business Networks (24)
Computer Systems (28)
Food Webs (7)
Internet (30)
Knowledge Networks (105)
Multi-Domain Representation (59)
Music (32)
Others (55)
Pattern Recognition (24)
Political Networks (20)
Semantic Networks (30)
Social Networks (89)
Transportation Networks (45)
World Wide Web (54)
See All (714)
VC Book is now in progress
VisualComplexity.com takes many hours of research and curation. You can support the
Display a menu

Radial Layouts

Uber Rides by Neighborhood

January 9, 2012

Mouseover to focus on rides to or from a single neighborhood.

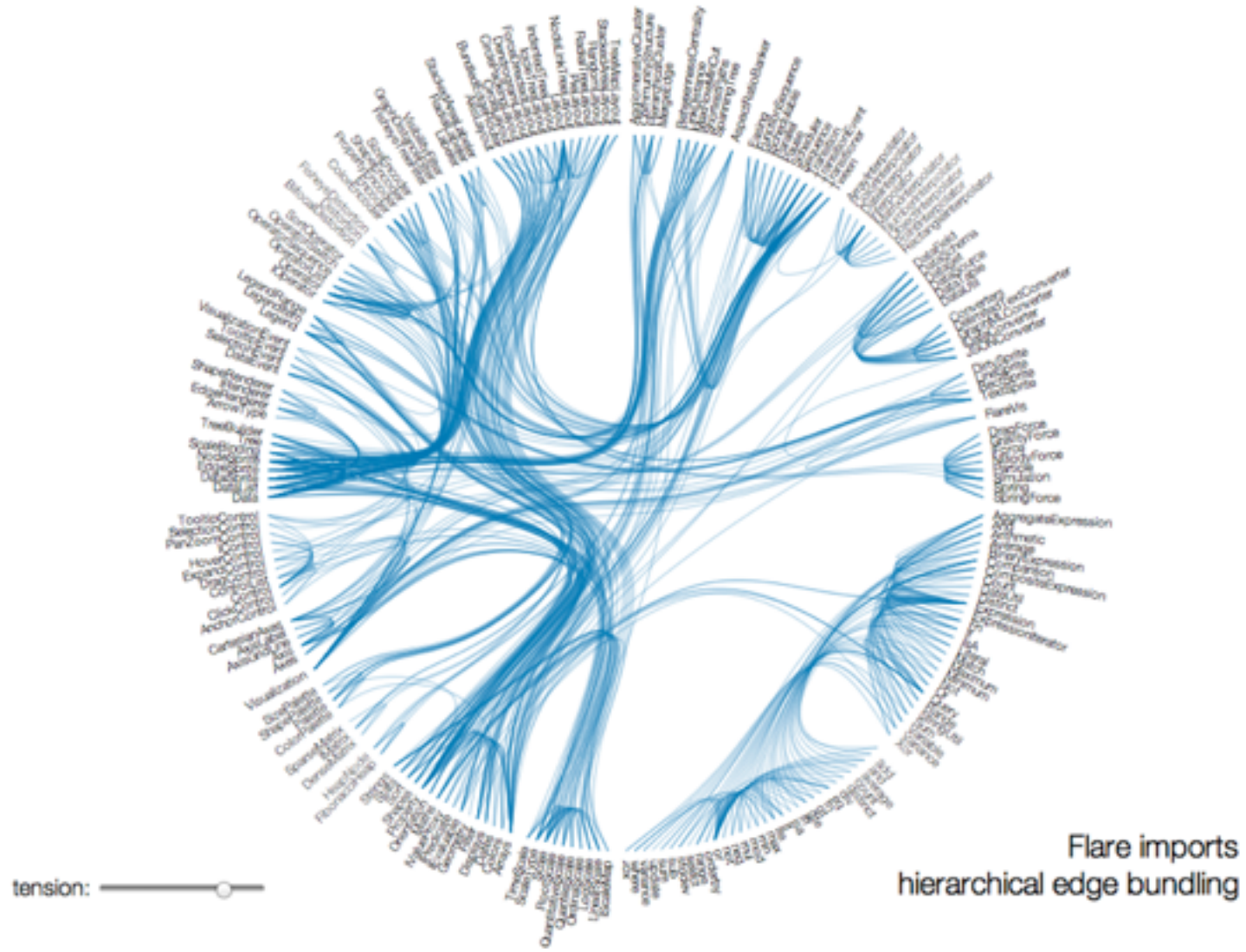
The thickness of links between neighborhoods encodes the relative frequency of rides between two neighborhoods: thicker links represent more frequent rides.

Links are directed: for example, while 2.2% of rides go from South of Market to Downtown, only 1.2% go in the opposite direction. Links are colored by the more frequent origin.

Scroll down for more!

Built with d3.js.

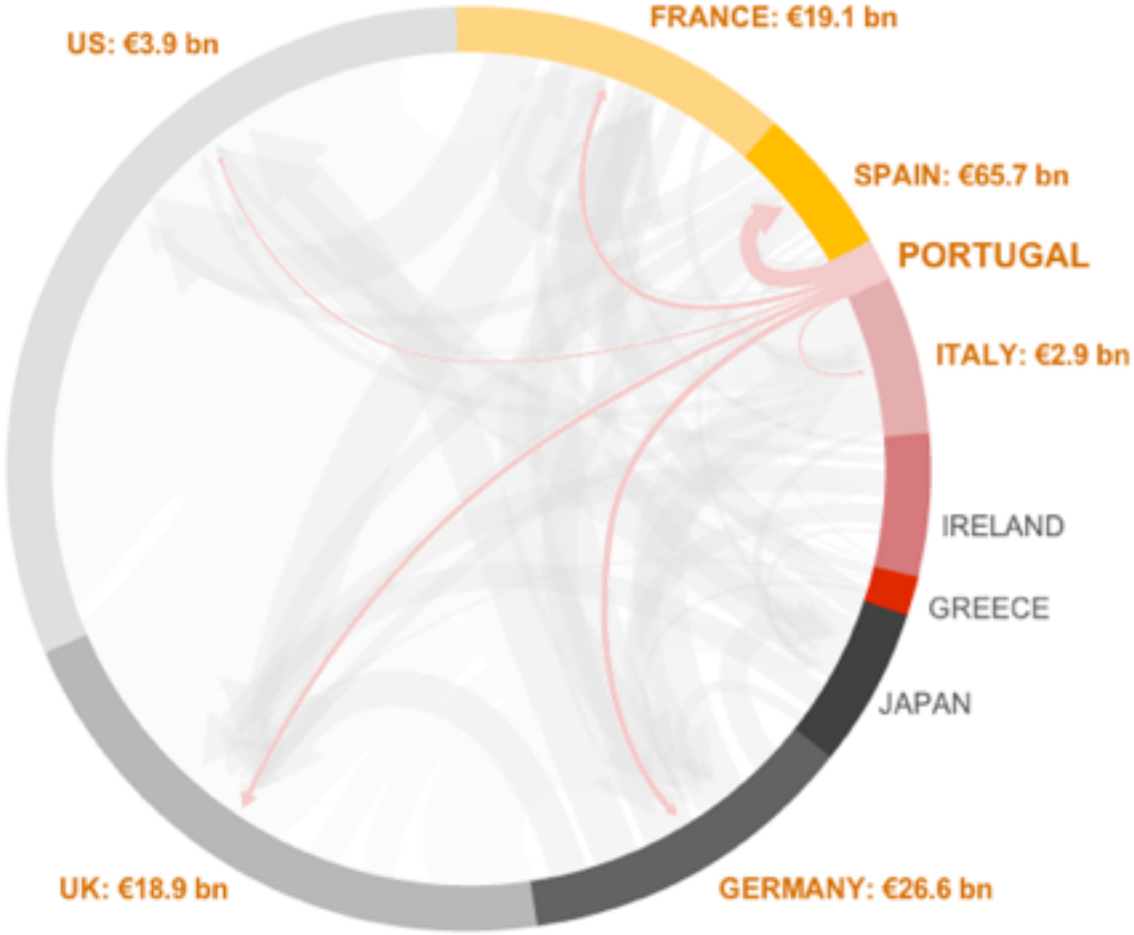
Hierarchical Edge Bundles



Michael Bostock

Eurozone debt web: Who owes what to whom?

The circle below shows the gross external, or foreign, debt of some of the main players in the eurozone as well as other big world economies. The arrows show how much money is owed by each country to banks in other nations. The arrows point from the debtor to the creditor and are proportional to the money owed as of the end of June 2011. The colours attributed to countries are a rough guide to how much trouble each economy is in.



PORTUGAL

GDP: €0.2 tn

Foreign debt: €0.4 tn

€38,081
Foreign debt per person

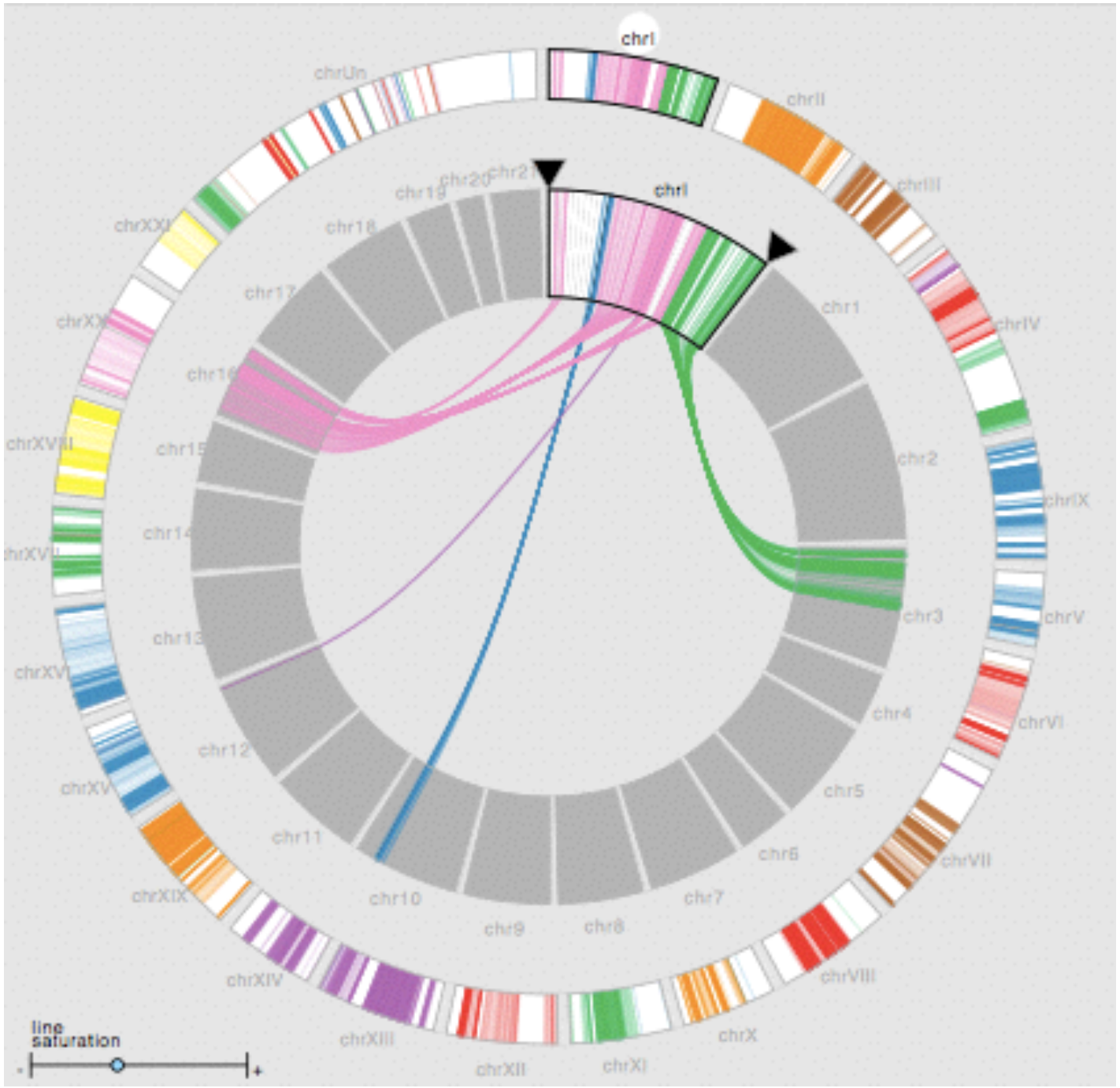
251%
Foreign debt to GDP

106%
Govt debt to GDP

Risk Status: HIGH

Portugal, the third eurozone country to need a bail-out, is in deep recession. It is currently implementing a series of austerity measures as well as planning a series of privatisations to fix its shaky finances and reduce its debt burden. The country is highly indebted to Spain, and its banks are owed 7.5bn euros by Greece.

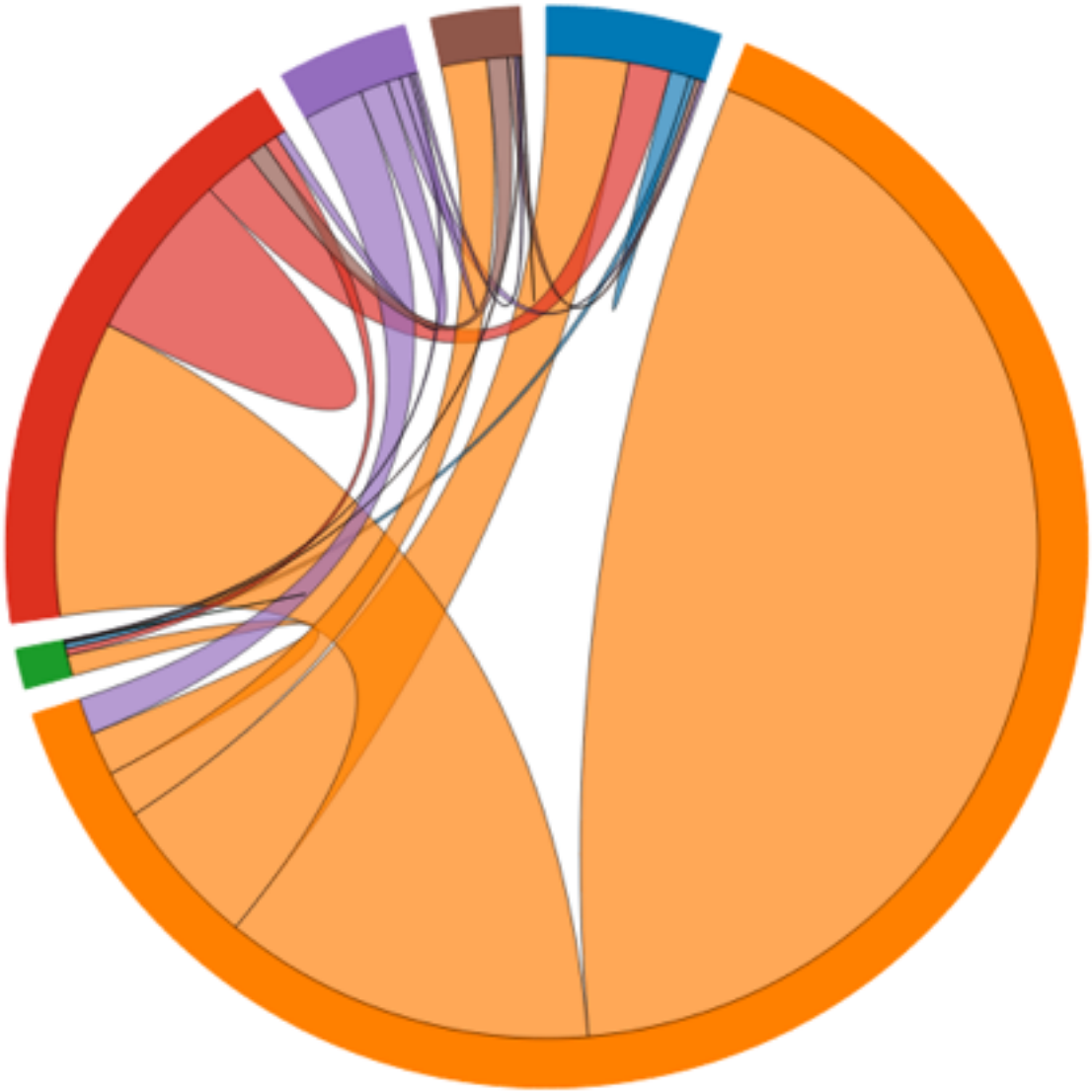
Back to introduction



MizBee, Meyer et al., 2009

Computed Relations Between Nodes as Chord Diagram

Nodes shown represent a high-scoring subset of the full 1644 node dataset from Moritz Stefaner's crawl of infovis tweeters in Summer 2011.



Select Data Set:

Small Eigen-Centric Subset

Nodes: 102

Edges: 3631

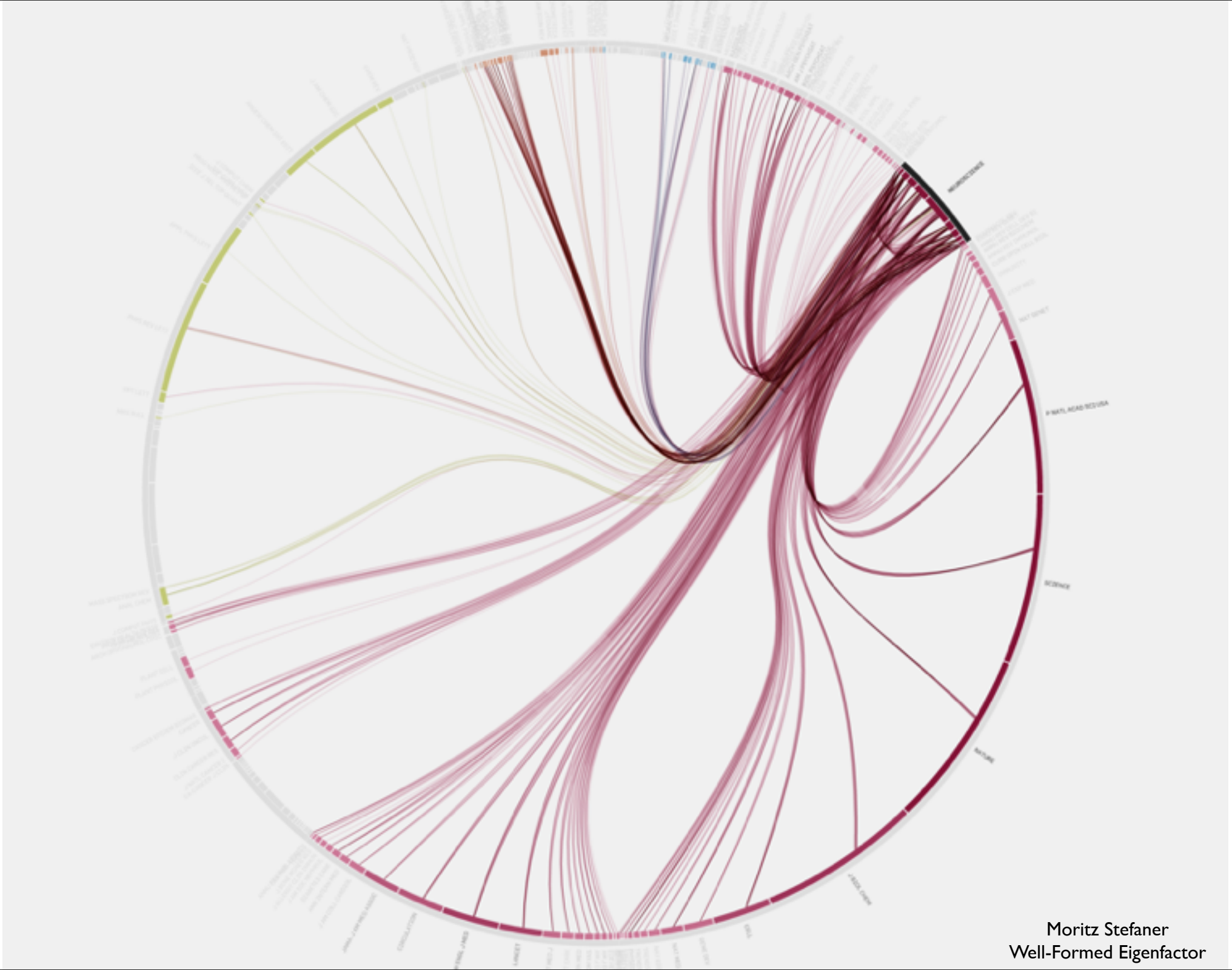
Computed Source/Target Matrix

	0	1	2	3	4	5
0	25	102	8	52	5	7
1	188	1626	55	467	45	58
2	7	26	1	8	1	2
3	64	370	18	210	12	24
4	17	75	5	35	12	5
5	9	57	3	26	4	2

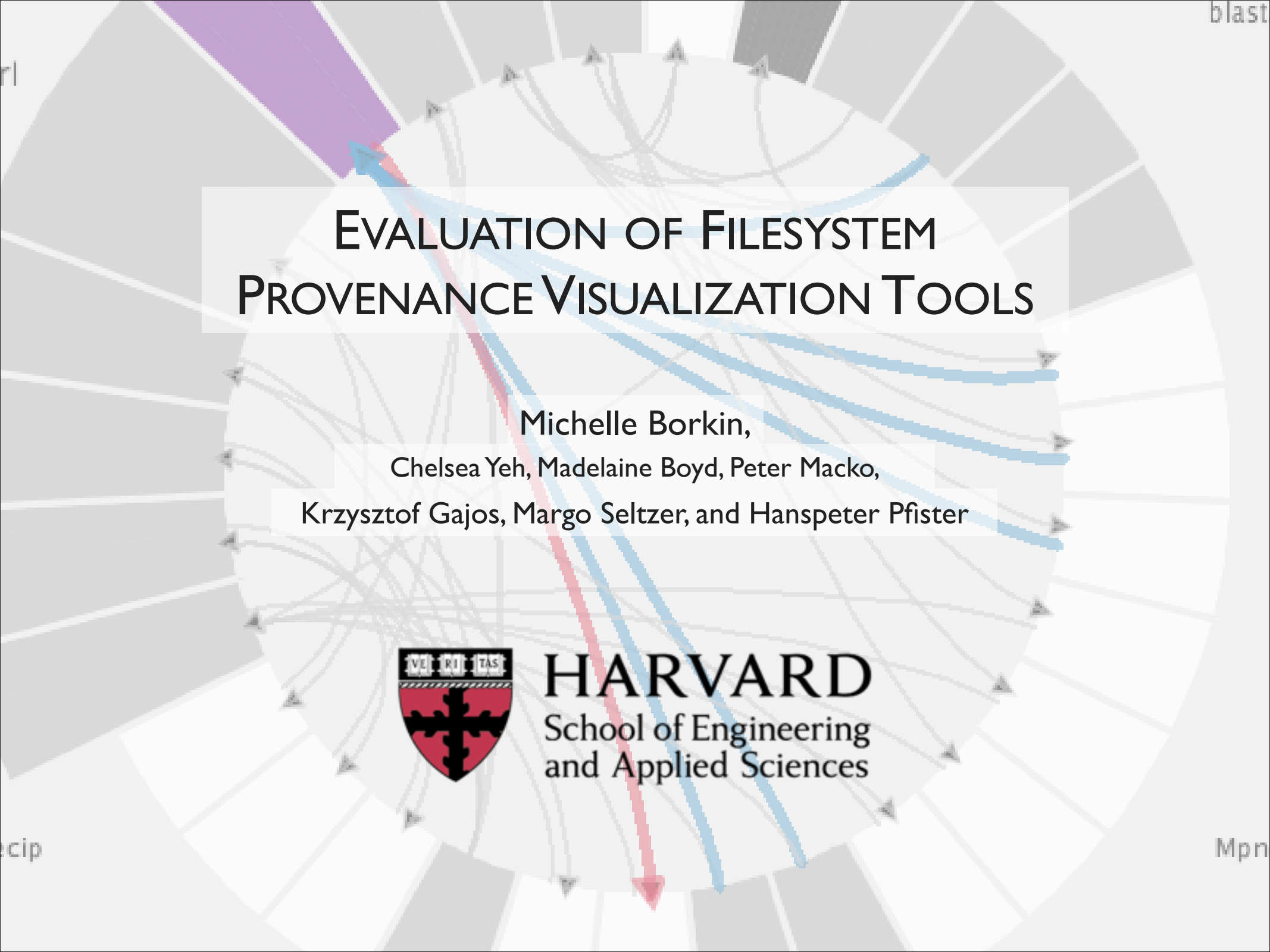
Color chords by larger:

(•) Source ( ) Target

Built with D3 by Lynn Cherny from NetworkX analysis with accompanying talk slides and blog post.



Moritz Stefaner
Well-Formed Eigenfactor



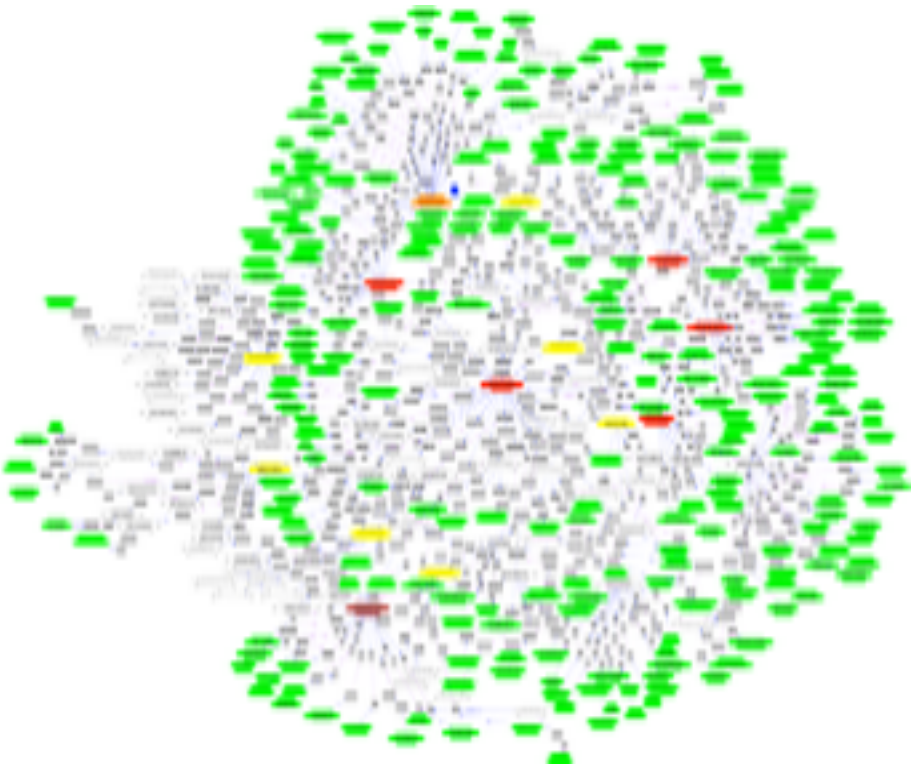
Evaluation of Filesystem Provenance Visualization Tools

Michelle Borkin,

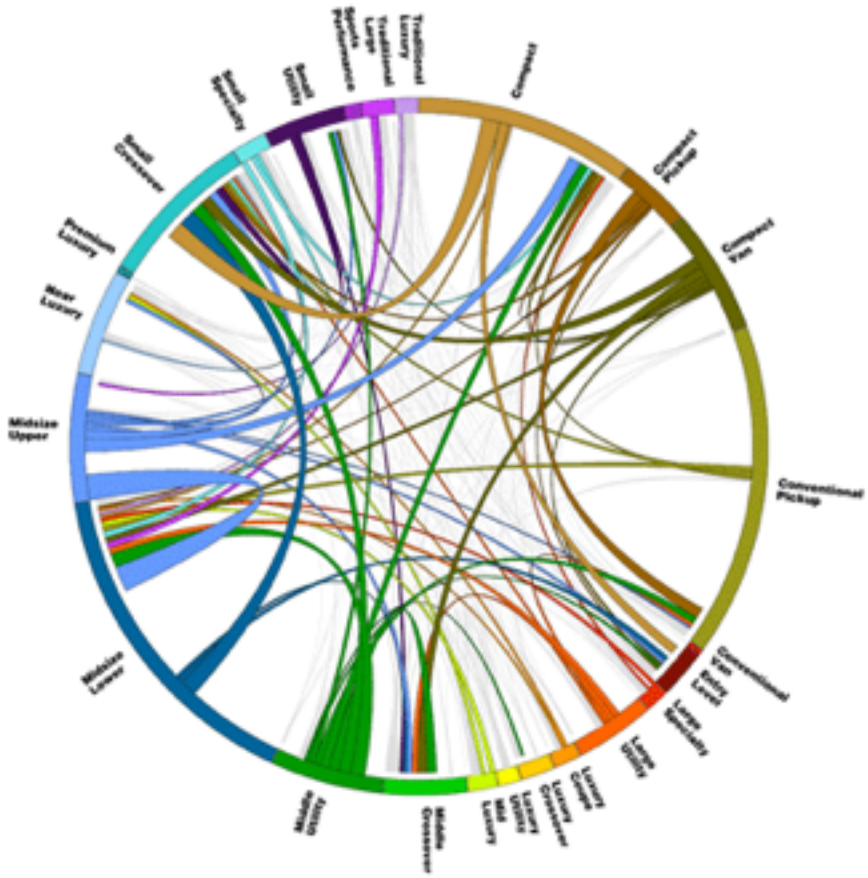
Chelsea Yeh, Madelaine Boyd, Peter Macko,

Krzysztof Gajos, Margo Seltzer, and Hanspeter Pfister

HARVARD
School of Engineering
and Applied Sciences



vs.



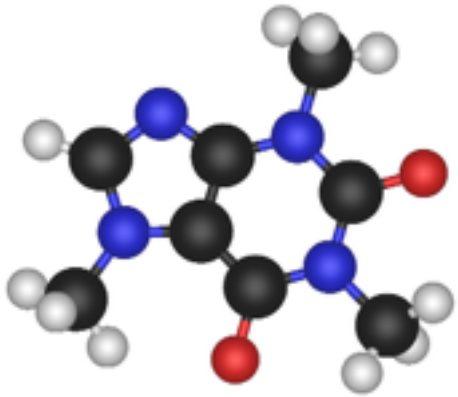
(Graphviz)

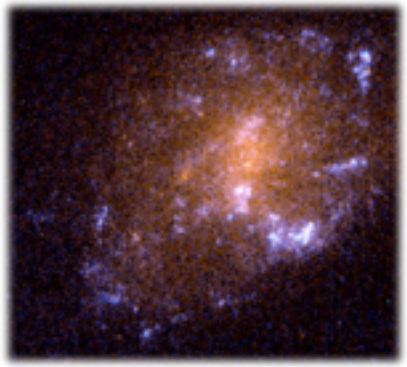
(Circos)

Borkin et al. "Evaluation of Filesystem Provenance Visualization Tools", VIS 2013

PROVENANCE DATA

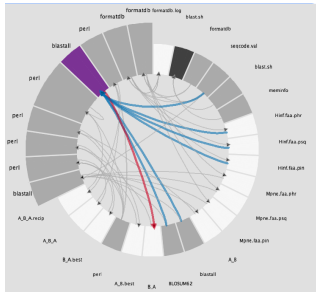
"A recording of the relationships of reads and writes between processes and files."



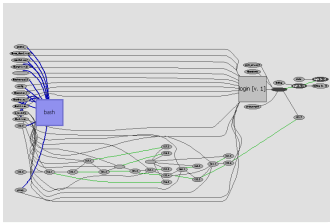


Borkin et al. "Evaluation of Filesystem Provenance Visualization Tools", VIS 2013

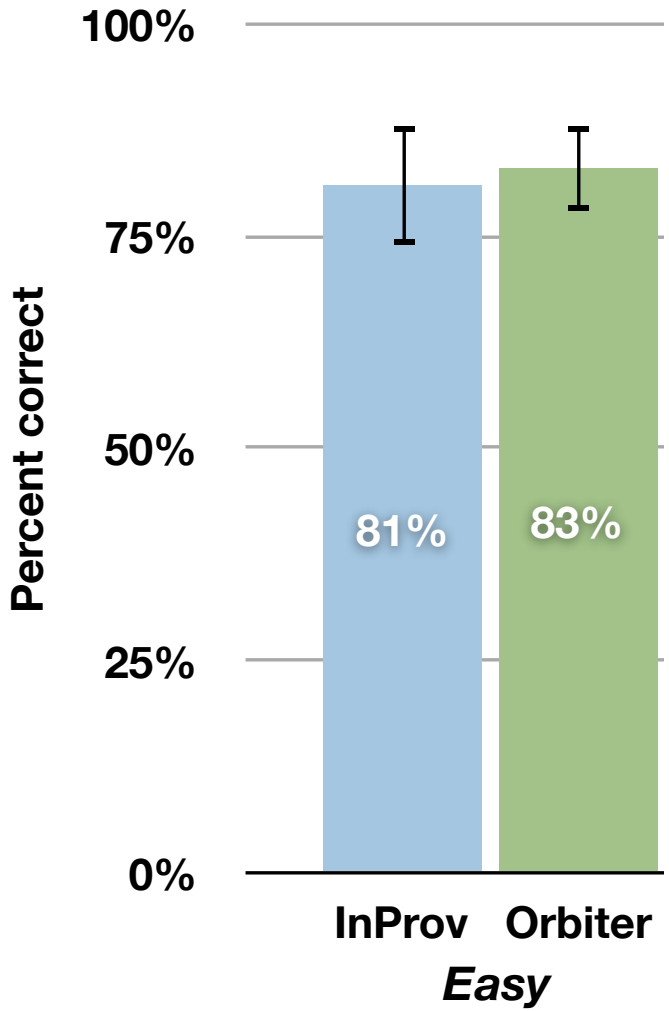
Accuracy



Radial
InProv

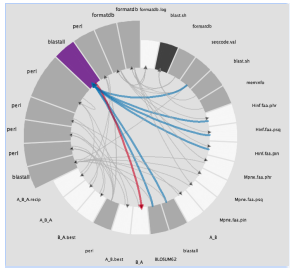


Node-link
Orbiter

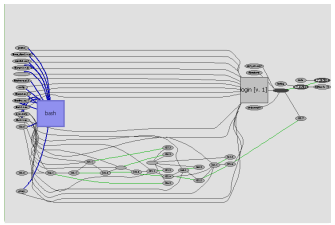


Borkin et al. "Evaluation of Filesystem Provenance Visualization Tools", VIS 2013

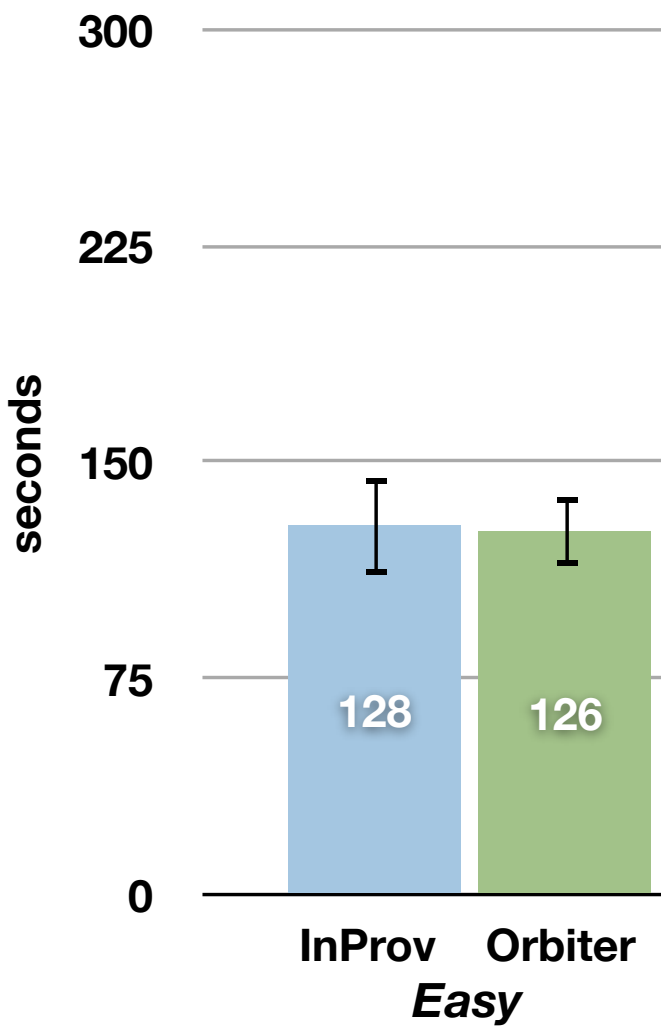
Efficiency



Radial



Node-link



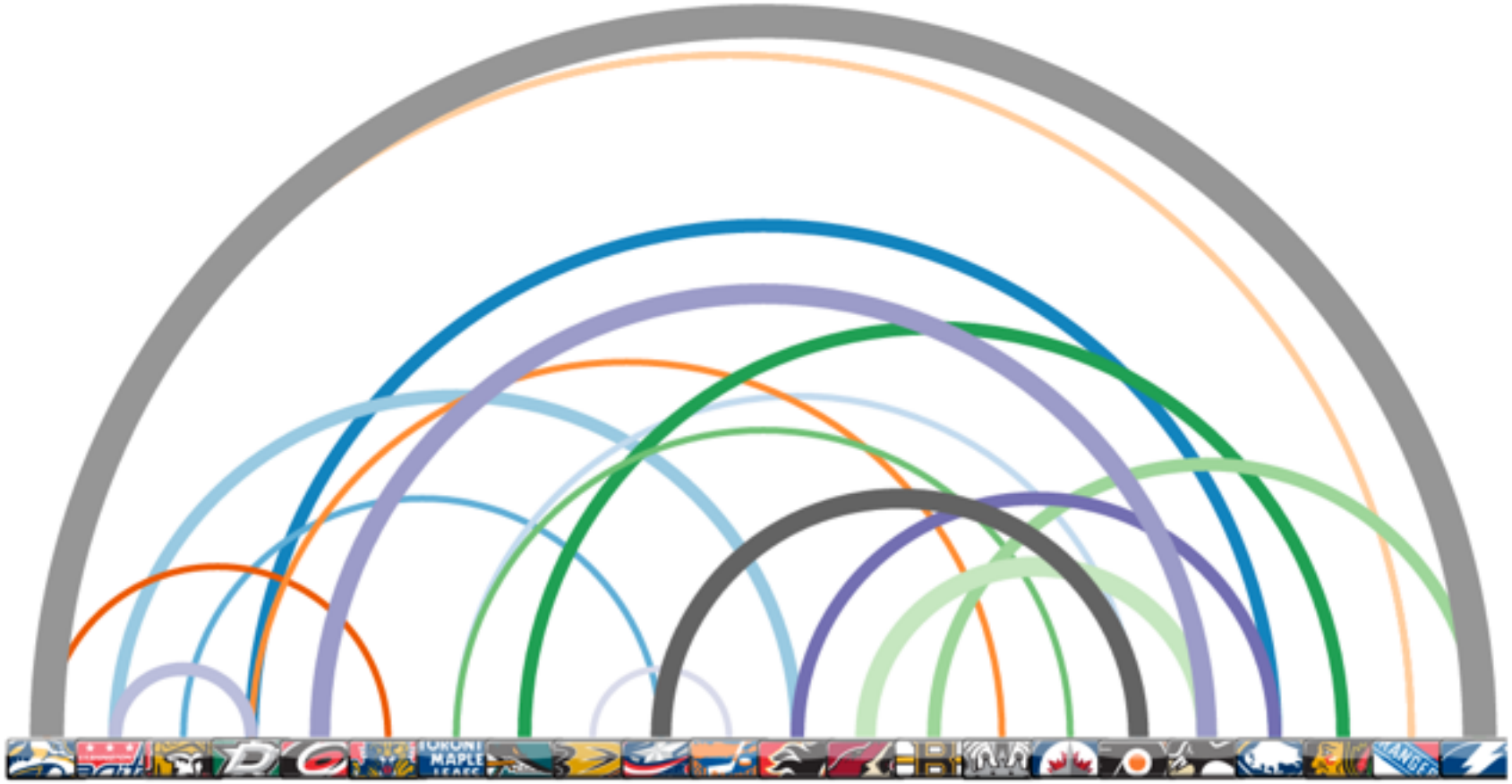
Borkin et al. "Evaluation of Filesystem Provenance Visualization Tools", VIS 2013

Linear Layouts

TradeArc

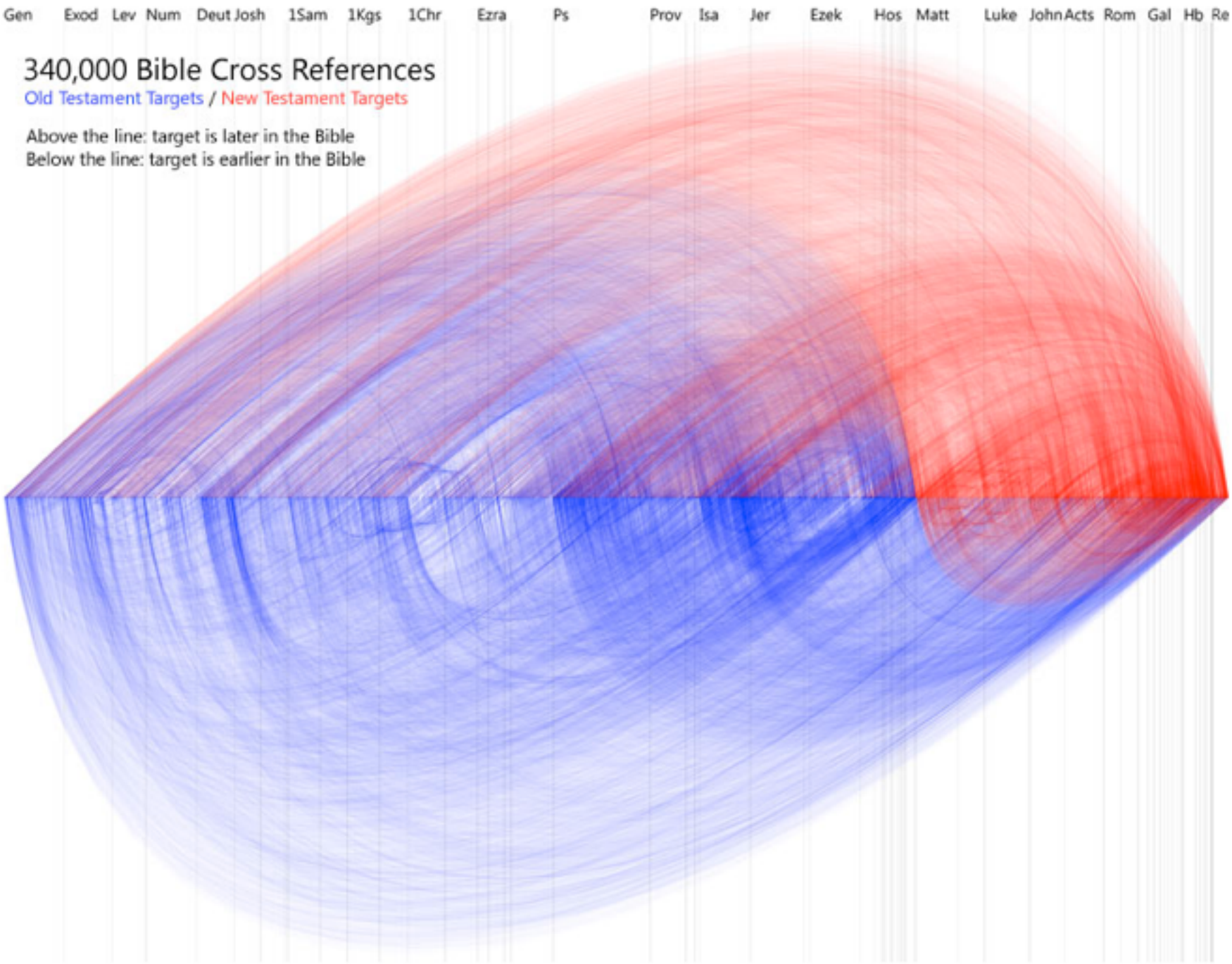
NHL trades that have taken place since June 15 2012

Click arc for trade details, or hover over a team to highlight its trades.

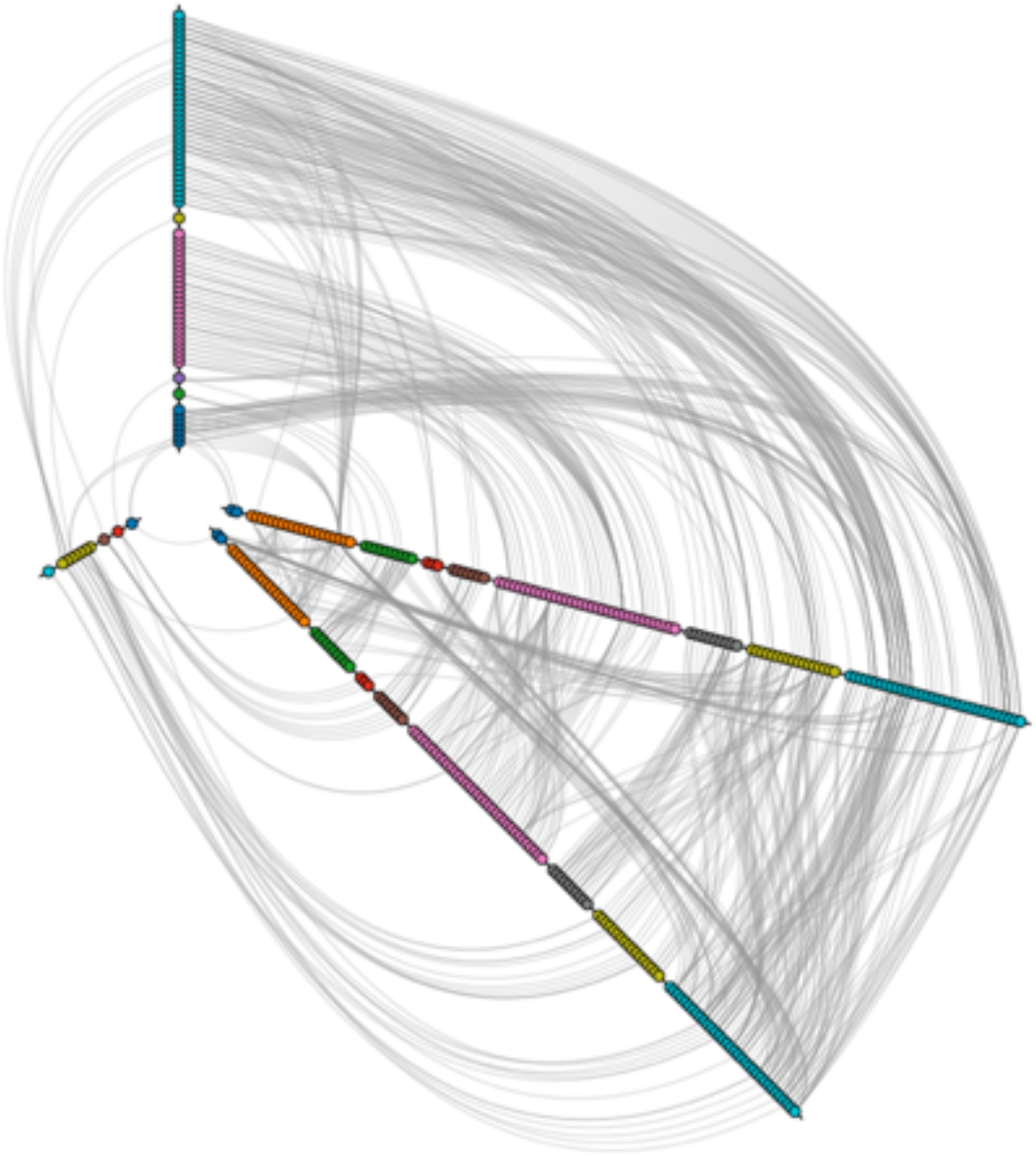


Bible Cross References Visualization

Here's a visualization of 340,000 Bible cross references:



Hive Plots

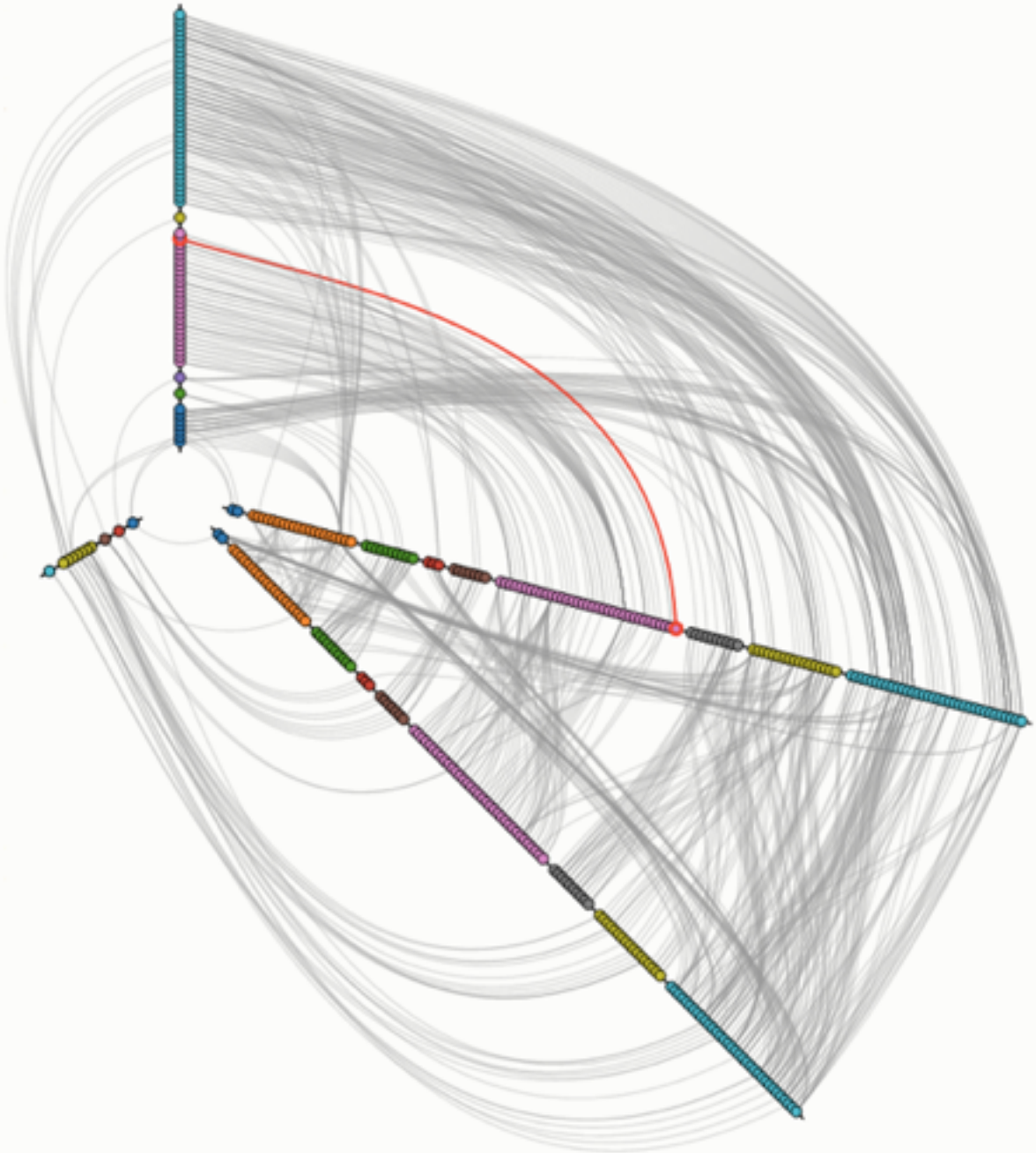


Martin Krzywinski
Michael Bostock

Hive Plots

March 18, 2012

flare.query.methods.xor → flare.query.Xor



This **hive plot**, a type of node-link diagram, shows the dependency graph of the Flare visualization toolkit.

Each dot represents a class, and each line represents an import statement from one class to another. Related classes share the same color.

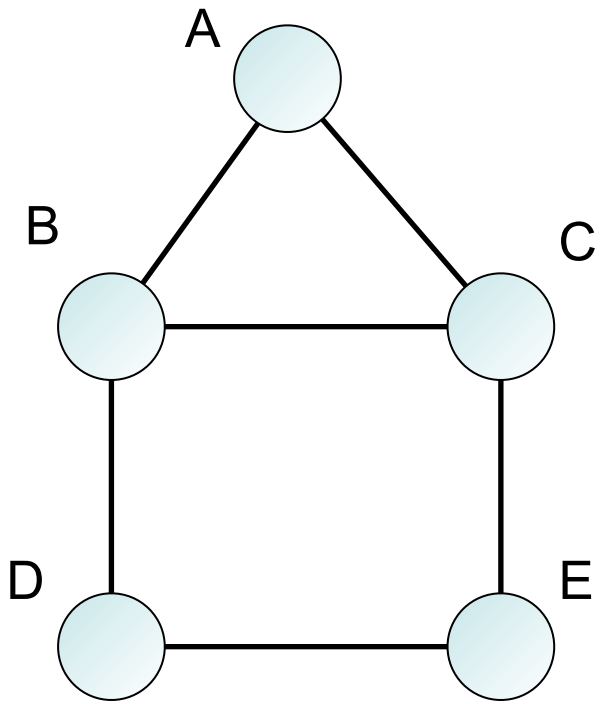
Mouseover for more details, and scroll down to read more about the use of hive plots to visualize networks.

Built with d3.js.

Matrices

Matrix layouts

Instead of node link diagram, use adjacency matrix representation



	A	B	C	D	E
A		■	■		
B	■		■	■	
C	■	■			■
D		■			■
E			■	■	

Spotting Patterns

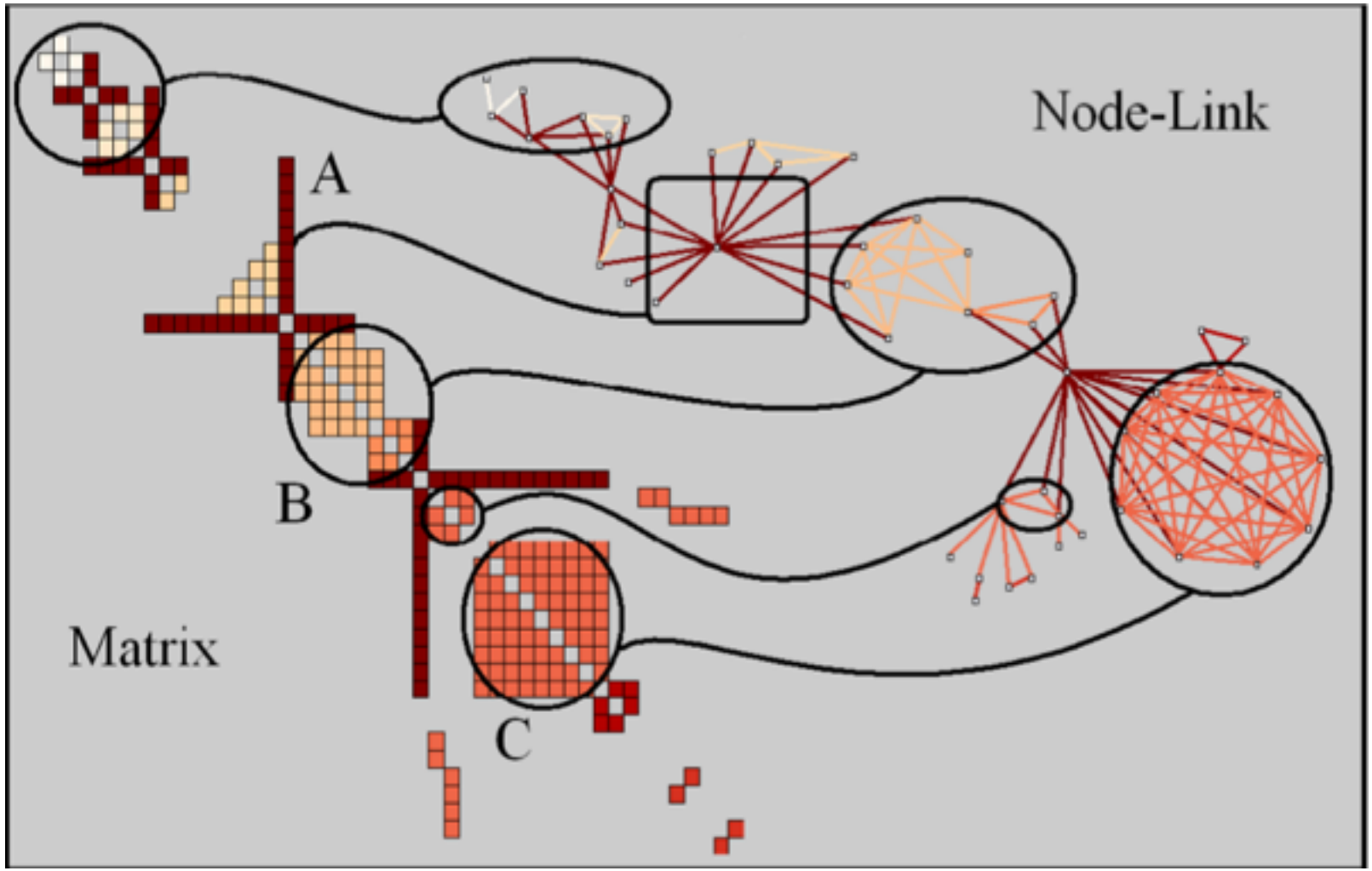
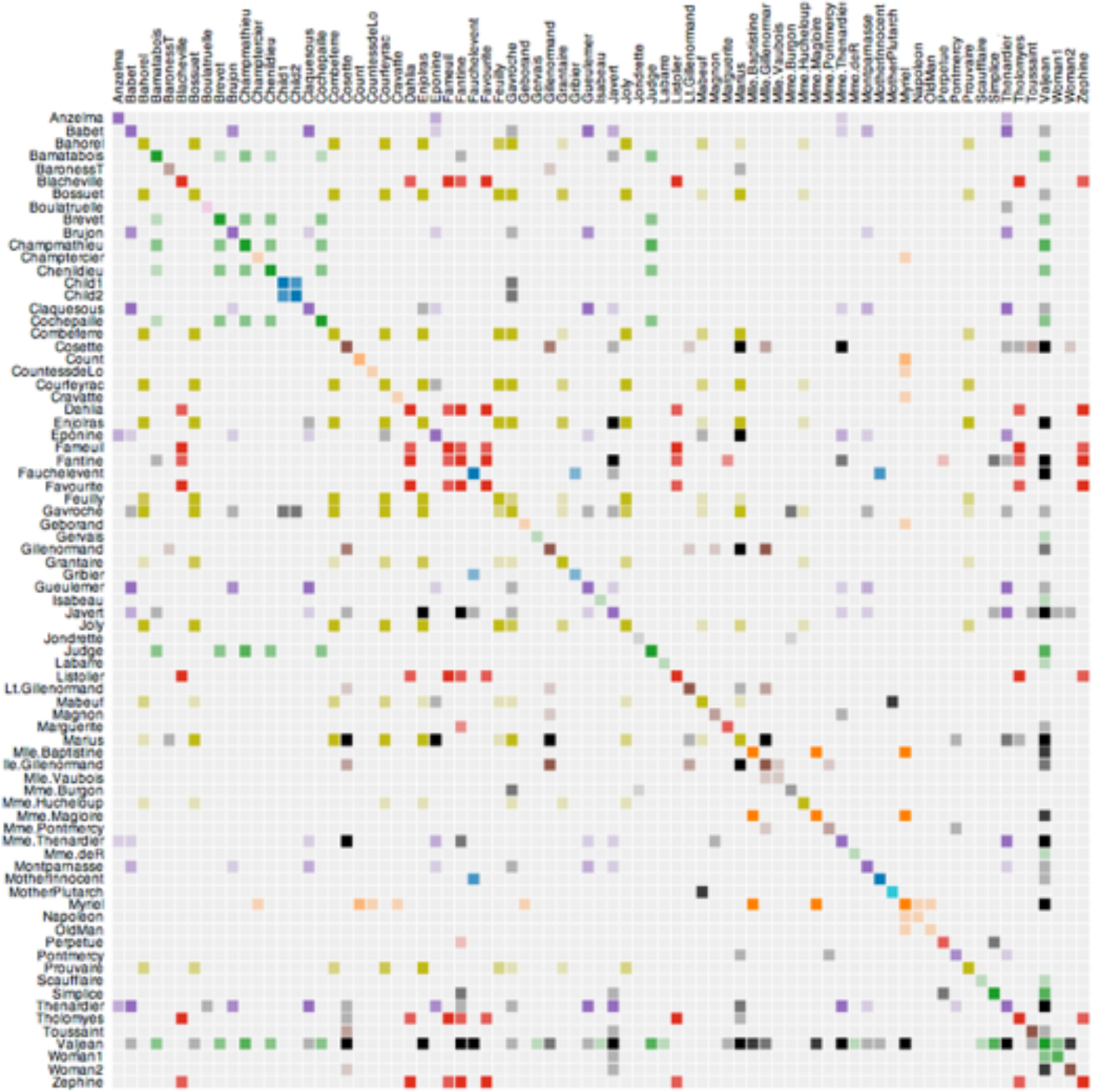
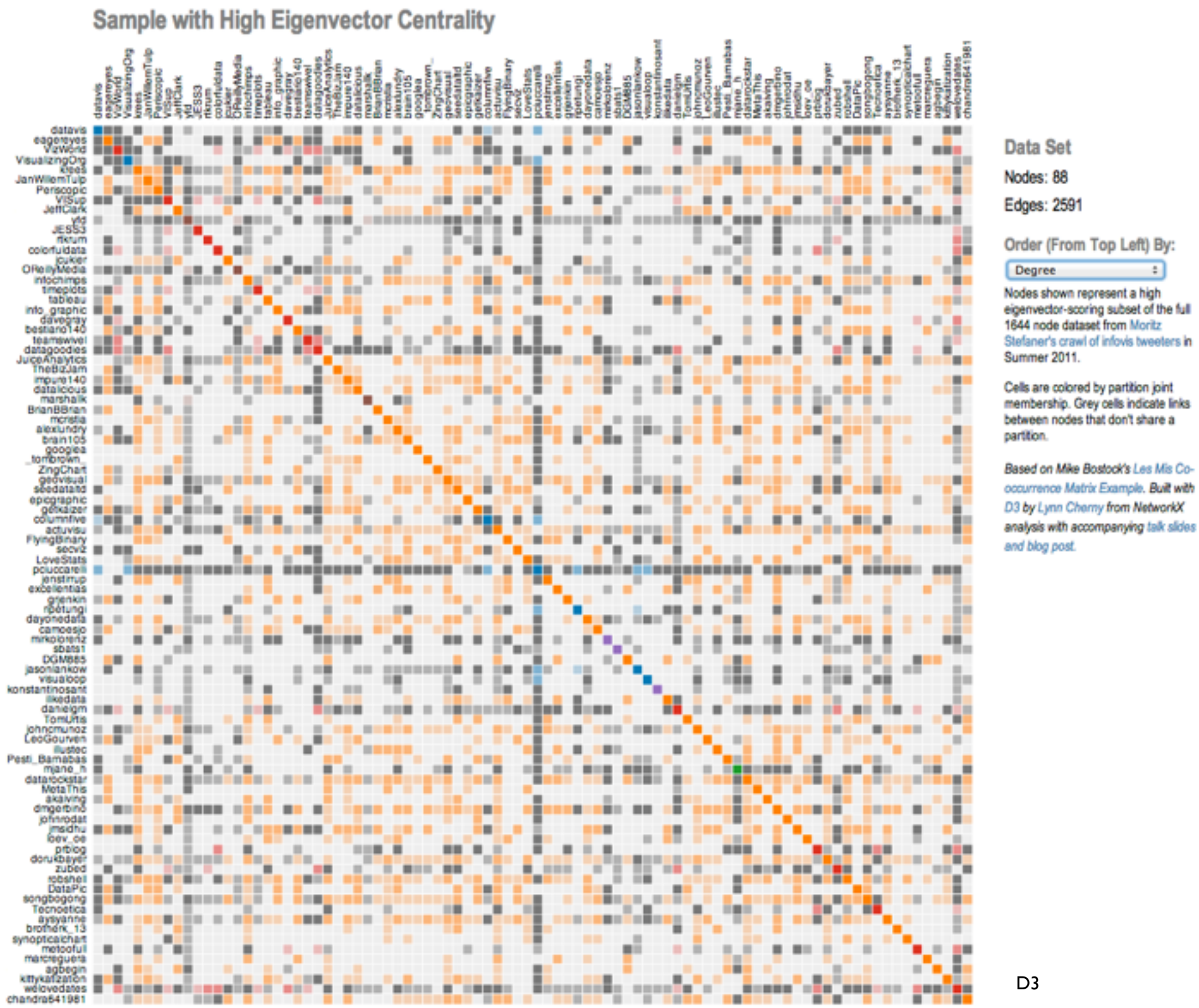


Image taken from : N. Henry and J.-D. Fekete MatrixExplorer: a Dual-Representation System to Explore Social Networks

Les Misérables Co-occurrence



Michael Bostock, D3


Sample with High Eigenvector Centrality
Data Set
Nodes: 88
Edges: 2591
Order (From Top Left) By:
Degree
Nodes shown represent a high eigenvector-scoring subset of the full 1644 node dataset from Moritz Stefaner's crawl of infovis tweeters in Summer 2011.
Cells are colored by partition joint membership. Grey cells indicate links between nodes that don't share a partition.
Based on Mike Bostock's Les Mis Co-occurrence Matrix Example. Built with D3 by Lynn Cherny from NetworkX analysis with accompanying talk slides and blog post.
D3

NodeTrix

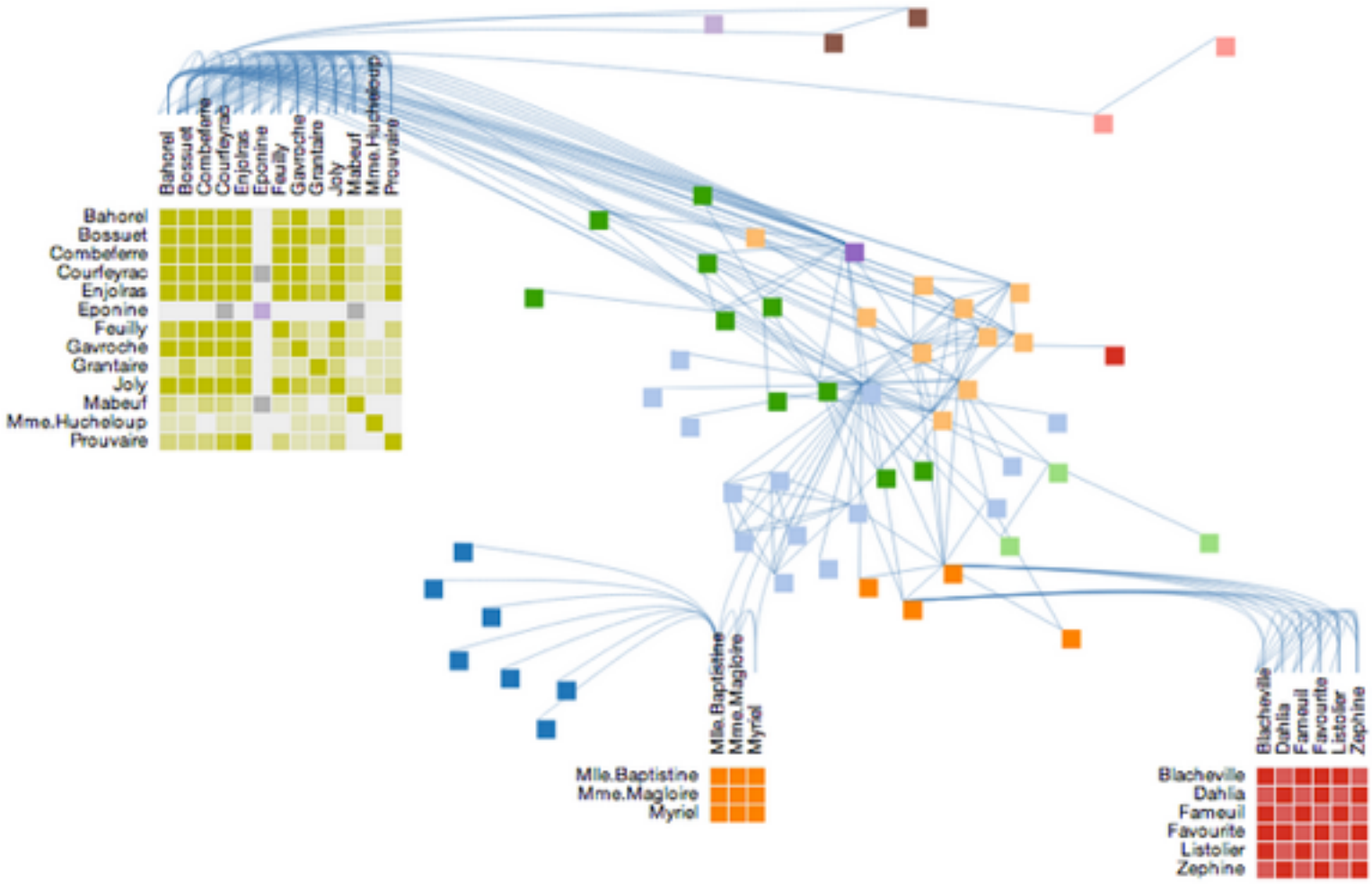
Research work

NodeTrix is a Node-Link Diagram hybrid visualization with Adjacency Matrix. Learn more about the research behind the tool.

Javascript version

Status: In progress

Below is a screenshot/teaser of the upcoming project:



Tools & Applications

Gephi
makes graphs handy

Download | Blog | Wiki | Forum | Support | Bug tracker

Home | Features | Plugins | Users | Developers | Partners

The Open Graph Viz Platform

Gephi is a visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

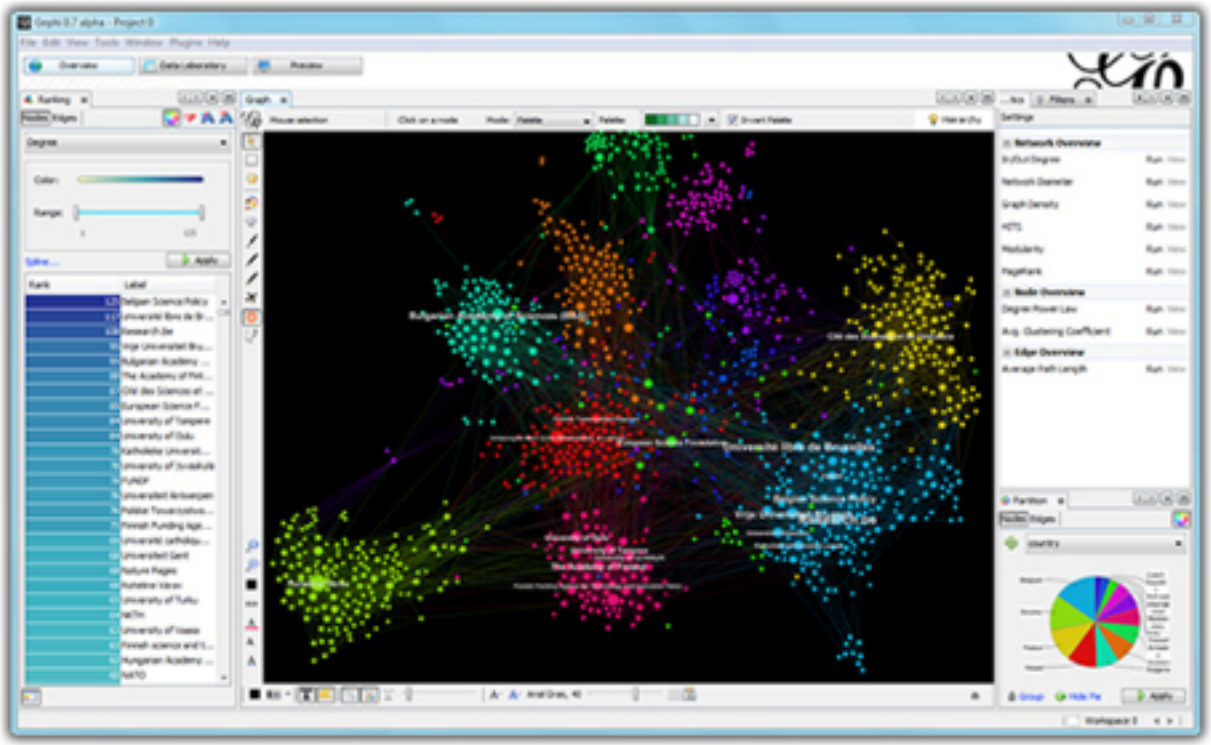
Runs on Windows, Linux and Mac OS X. Gephi is open-source and free.

Learn More on Gephi Platform »

Download FREE
Gephi 0.7 alpha

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos



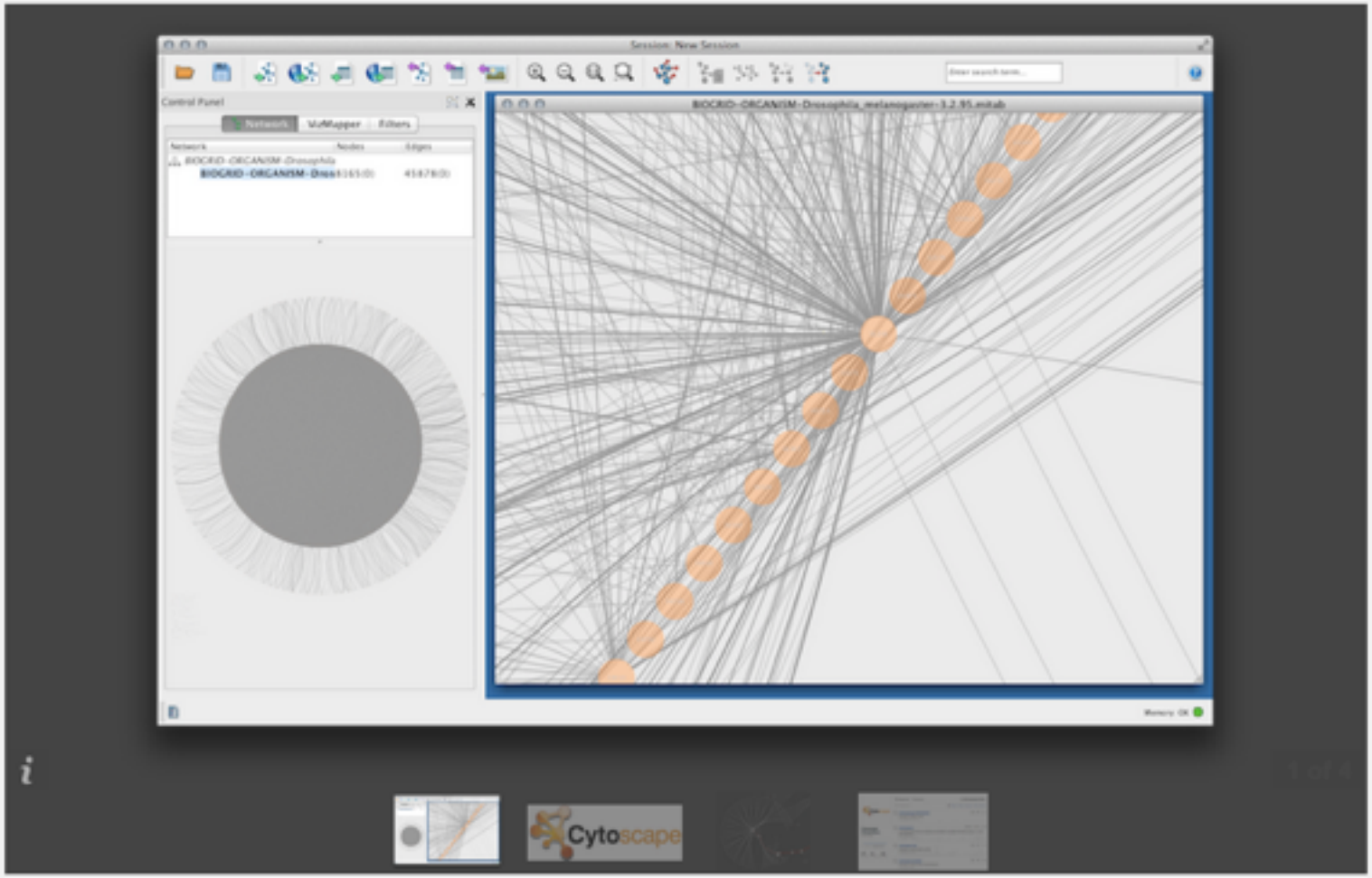
Gephi has been accepted again for Google Summer of Code! The program is the best way for students around the world to start contributing to an open-source project. Students, apply now for Gephi proposals. Come to the GSOC forum section and say Hi! to this topic.

Learn More »

Cytoscape

Search... Go

Home | Introduction | Download | Apps | Documentation | Community | Report a Bug | Getting Help



Network Data Integration, Analysis, and Visualization in a Box

Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. A lot of *Apps* are available for various kinds of problem domains, including bioinformatics, social network analysis, and semantic web.

Download Cytoscape

Welcome Letter

Release Notes

Sample Visualizations

Learning Cytoscape

Introduction to Cytoscape | Tutorial for 3.x | Tutorial for 2.x | Developers Tutorial

NetworkX

High-productivity software for complex networks

NetworkX is a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

Documentation
all documentation

Examples
using the library

Reference
all functions and methods

Features

- Python language data structures for graphs, digraphs, and multigraphs.
- Nodes can be "anything" (e.g. text, images, XML records)
- Edges can hold arbitrary data (e.g. weights, time-series)
- Generators for classic graphs, random graphs, and synthetic networks
- Standard graph algorithms
- Network structure and analysis measures
- Open source BSD license
- Well tested: more than 1800 unit tests, >90% code coverage
- Additional benefits from Python: fast prototyping, easy to teach, multi-platform

Versions

Latest Release

1.8.1 - 4 August 2013
downloads | docs | pdf

Development

1.9dev
github | docs | pdf

Contact

Mailing list
Issue tracker
Developer guide

Circos is back for 3rd year at 2012 Bioinformatics and Comparative Genome Analysis course by the Pasteur Institute—May 9

Google™ Custom Search Search ×

WHAT IS CIRCOS?

CIRCULAR VISUALIZATION

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships

The igraph library

sourceforge

Home | News | Download | Documentation | Wiki | Screenshots | Mailing lists | Bugs | License

Introduction

igraph is a free software package for creating and manipulating undirected and directed graphs. It includes implementations for classic graph theory problems like minimum spanning trees and network flow, and also implements algorithms for some recent network analysis methods, like community structure search.

Read more »

Features

igraph contains functions for generating regular and random graphs, manipulating graphs, assigning attributes to vertices and edges. It can calculate various structural properties, graph isomorphism, includes heuristics for community structure detection, supports many file formats. The R and Python interfaces support visualization.

Read more »

Requirements

igraph runs on most modern machines and operating systems, and it is tested on MS Windows, Mac OSX and various Linux versions.

The software you need for installing **igraph** depends on whether you want to use the C library, the R package or the Python extension; and may vary depending on your platform.

Read more »

Latest version: **0.6.5**
Release notes

Latest Announcements

igraph 0.6.5 *(3 Mar 2013)*

igraph 0.6.5 was officially released today.

This is a bugfix release, but it has some nice new features as well, see some of them in the release notes:
http://igraph.sourceforge.net/relnotes-0.6.5.html

Databases

Database

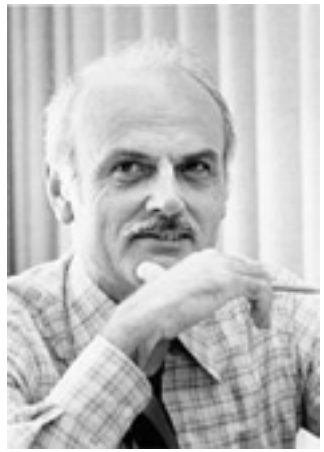
- Database: A large collection of *structured* data
- Database Management System (DBMS): Software that stores, manages, and facilitates access to databases
- Traditionally, relational databases with transactions
- Modern usage varies (NoSQL, maps, etc.)

Database

- Models a real-world enterprise
- Entities (e.g., teams, games)
- Relationships (e.g., Red Sox *play against* Cardinals *in the* World Series)
- Can also include "business logic" (e.g., the MLB ranking system)

Relational Model

"[The Relational Model] provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation on the other."



E.F. Codd, 1981 Turing Award winner

Key Concepts

- Data model: a collection of concepts for describing data
- Schema: a description of a particular collection of data, using a given data model
- Relational model:
  - Relation: table with rows and columns
  - Schema: describes columns, or fields

Example: Harvard SIS

- Conceptual Schema

  Students(sid: string, name: string, age: integer, gpa:real)
  Courses(cid: string, cname:string, credits:integer)
  Enrolled(sid:string, cid:string, grade:string)
  FOREIGN KEY sid REFERENCES Students
  FOREIGN KEY cid REFERENCES Courses

- External Schema (View)

  Course_info(cid:string,enrollment:integer)
  Create View Course_info AS
  SELECT cid, Count (*) as enrollment FROM Enrolled
  GROUP BY cid

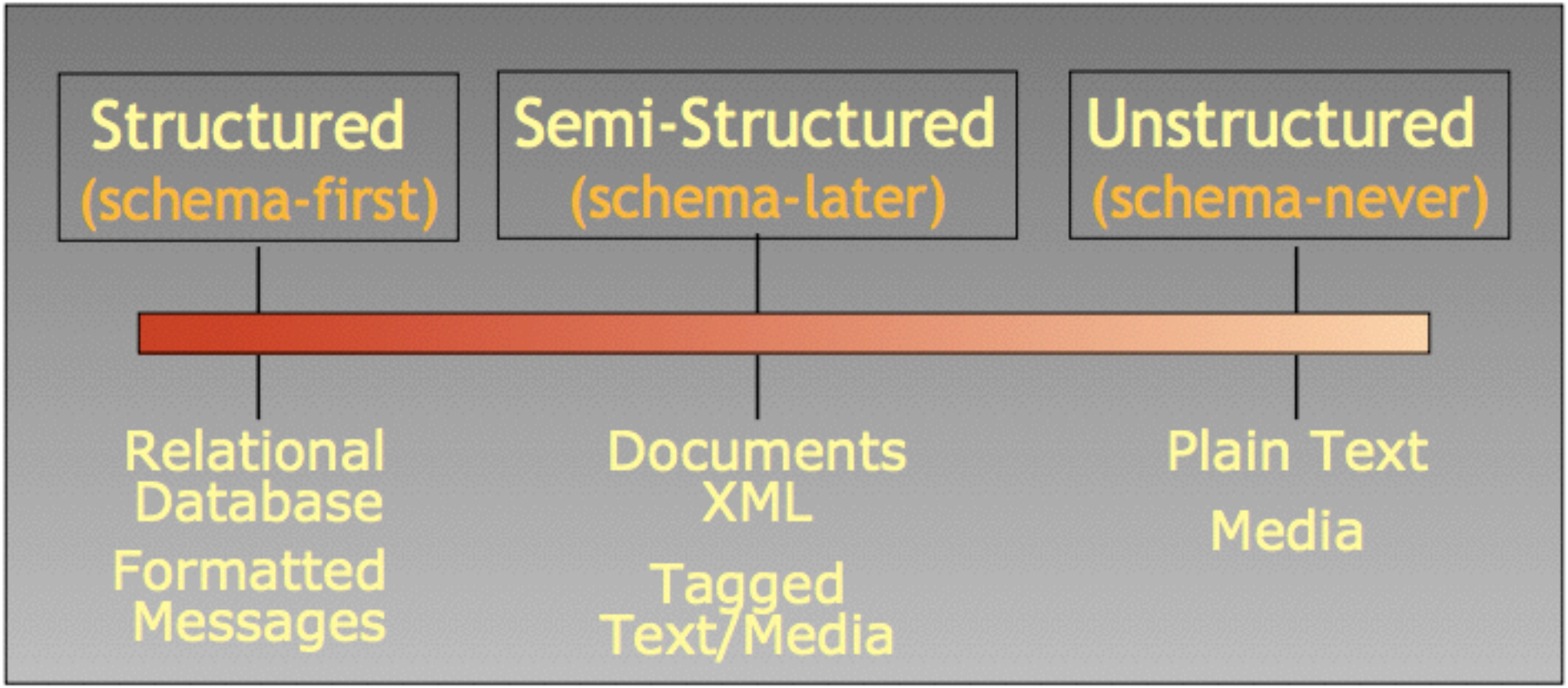
Instance of Students Relation

sid	name	login	age	gpa
7689234	Jones	jones@seas	18	3.4
7636112	Smith	smith@seas	19	3.3
7632483	Smith	smith@math	18	3.8

DBMS

- Stores, manages, and facilitates access to databases
- Provides:
  - Data Definition Language (DDL)
  - Data Manipulation Language (DML)
    - Queries - to retrieve analyze, and modify data
  - Guarantees about durability, concurrency, semantics, etc.

Structure Spectrum



Data Independence

"Relational DataBase Management Systems were invented to let you use one set of data in multiple ways, including ways that are unforeseen at the time the database is built and the 1st applications are written."

Curt Monash, Analyst / Blogger

Is a file system a DBMS?

- Thought experiment 1:

  You and your partner are editing the same file

  You both save it at the same time

  Whose changes survive?

  a) Yours b) Partner's c) Both d) Neither e) ???

- Thought experiment 2:

  You're updating a file and the power goes out

  Which of your changes survive?

  a) All b) None c) All since last saved d) ???

Is the WWW a DBMS?

- On the surface: documents and search

  Crawler *indexes* pages on the web

  Keyword-based *search* for pages

- Source data is mostly unstructured and untyped

  But more and more XML and JSON

- Public interface is search only

  Cannot modify data, no summaries, complex combinations, etc.

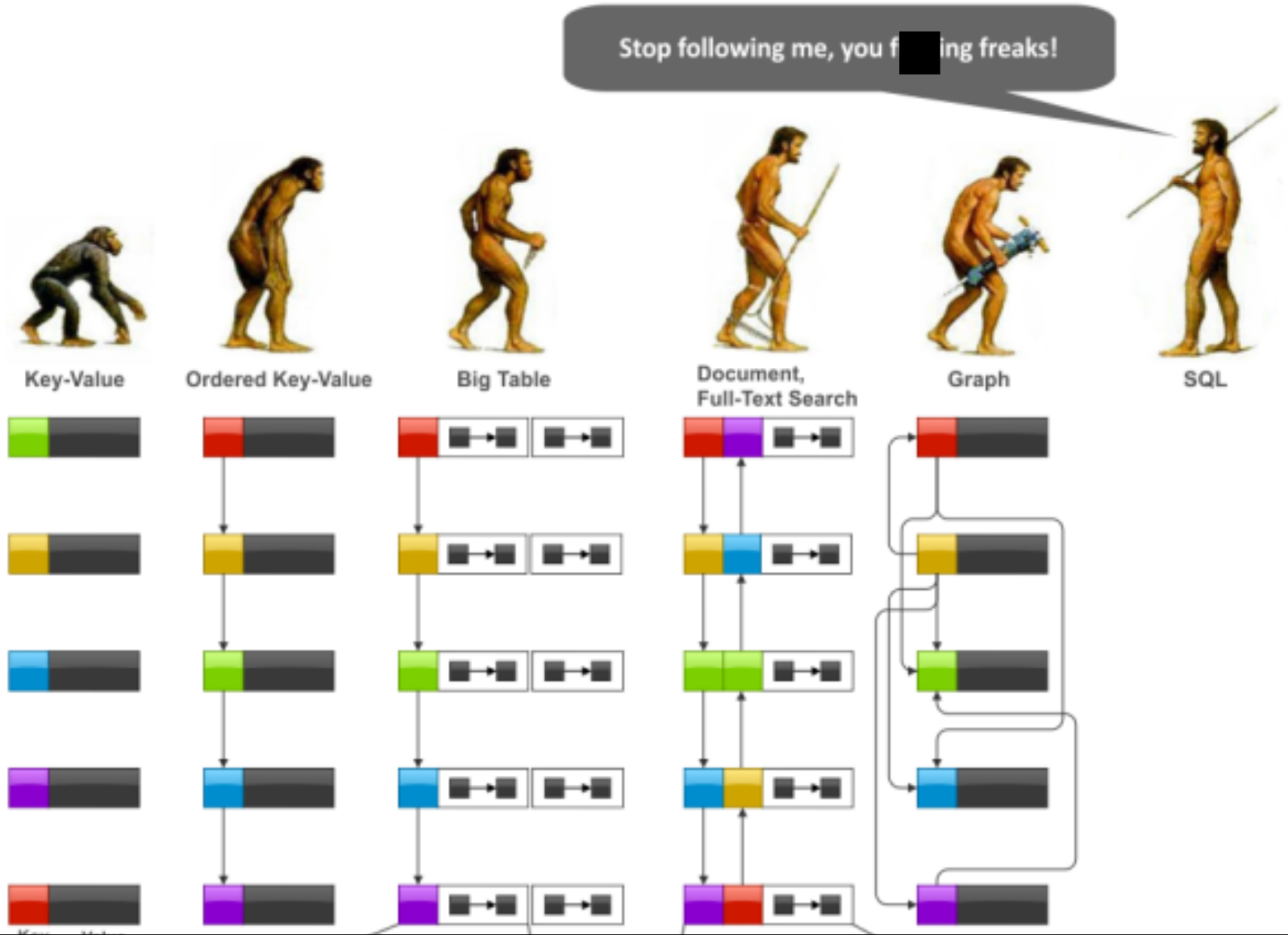
- Few guarantees for freshness, consistency, fault tolerance, ...

Current Market

- Relational DBMSs
  - Elephants: Oracle, IBM, Microsoft, Teradata, HP, EMC, ...
  - Open source: MySQL, PostgreSQL
- Search
  - Google & Bing
- Open Source "NoSQL"
  - Hadoop MapReduce
  - Key-value stores: Cassandra, Mongo, Riak, Voldemort, ...
- Cloud services
  - Amazon, Google AppEngine, MS Azure, Heroku, ...
- Increasing use of custom code

NoSQL Data Models

http://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/



AWS Reference Architectures

- Amazon EC2
- Amazon EC2 Spot
- Amazon EMR
- Amazon S3
- Amazon CloudFront
- Amazon RDS

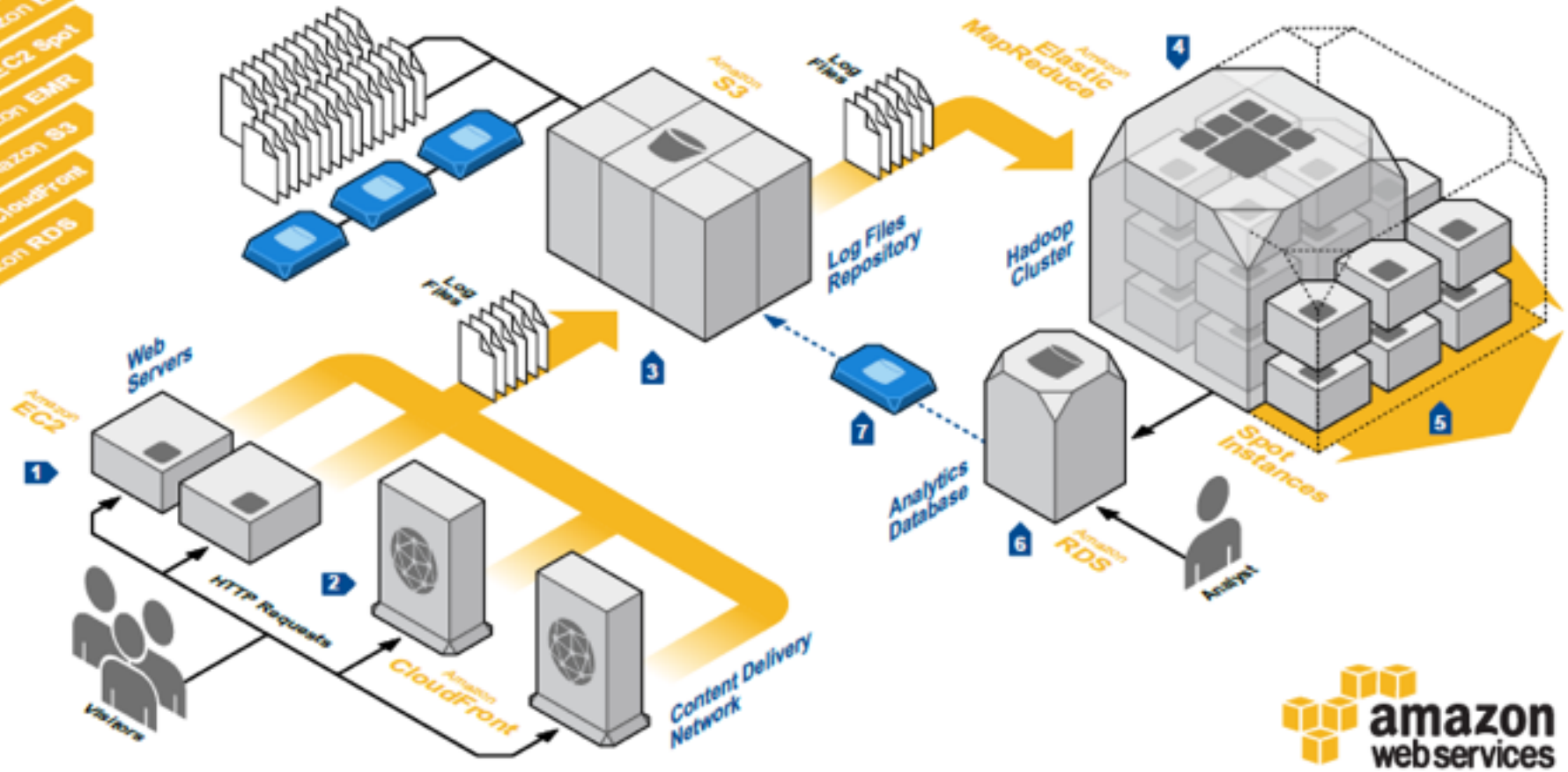
WEB LOG ANALYSIS

Amazon Web Services provides services and infrastructure to build reliable, fault-tolerant, and highly available web applications in the cloud. In production environments, these applications can generate huge amounts of log information.

This data can be an important source of knowledge for any company that is operating web applications. Analyzing logs can reveal information such as traffic patterns, user behavior, marketing profiles, etc.

However, as the web application grows and the number of visitors increases, storing and analyzing web logs becomes increasingly challenging.

This diagram shows how to use Amazon Web Services to build a scalable and reliable large-scale log analytics platform. The core component of this architecture is Amazon Elastic MapReduce, a web service that enables analysts to process large amounts of data easily and cost-effectively using a Hadoop hosted framework.



amazon web services

System Overview

1. The web front-end servers are running on **Amazon Elastic Compute Cloud** (Amazon EC2) instances.

2. **Amazon CloudFront** is a content delivery network that uses low latency and high data transfer speeds to distribute static files to customers. This service also generates valuable log information.

3. Log files are periodically uploaded to **Amazon Simple Storage Service** (Amazon S3), a highly available and reliable data store. Data is sent in parallel from multiple web servers or edge locations.

4. An **Amazon Elastic MapReduce** cluster processes the data set. **Amazon Elastic MapReduce** utilizes a hosted Hadoop framework, which processes the data in a parallel job flow.

5. When **Amazon EC2** has unused capacity, it offers EC2 instances at a reduced cost, called the *Spot Price*. This price fluctuates based on availability and demand. If your workload is flexible in terms of time of completion or required capacity, you can dynamically extend the capacity of your cluster using **Spot Instances** and significantly reduce the cost of running your job flows.

6. Data processing results are pushed back to a relational database using tools like *Apache Hive*. The database can be an **Amazon Relational Database Service** (Amazon RDS) instance. **Amazon RDS** makes it easy to set up, operate, and scale a relational database in the cloud.

7. Like many services, **Amazon RDS** instances are priced on a pay-as-you-go model. After analysis, the database can be backed-up into **Amazon S3** as a database snapshot, and then terminated. The database can then be recreated from the snapshot whenever needed.

Guest Lecture

- Margo Seltzer
- Herchel Smith Professor of Computer Science and a Harvard College Professor
- Tuesday, 10/29:
  Web Scale Data Management: An Historical Perspective