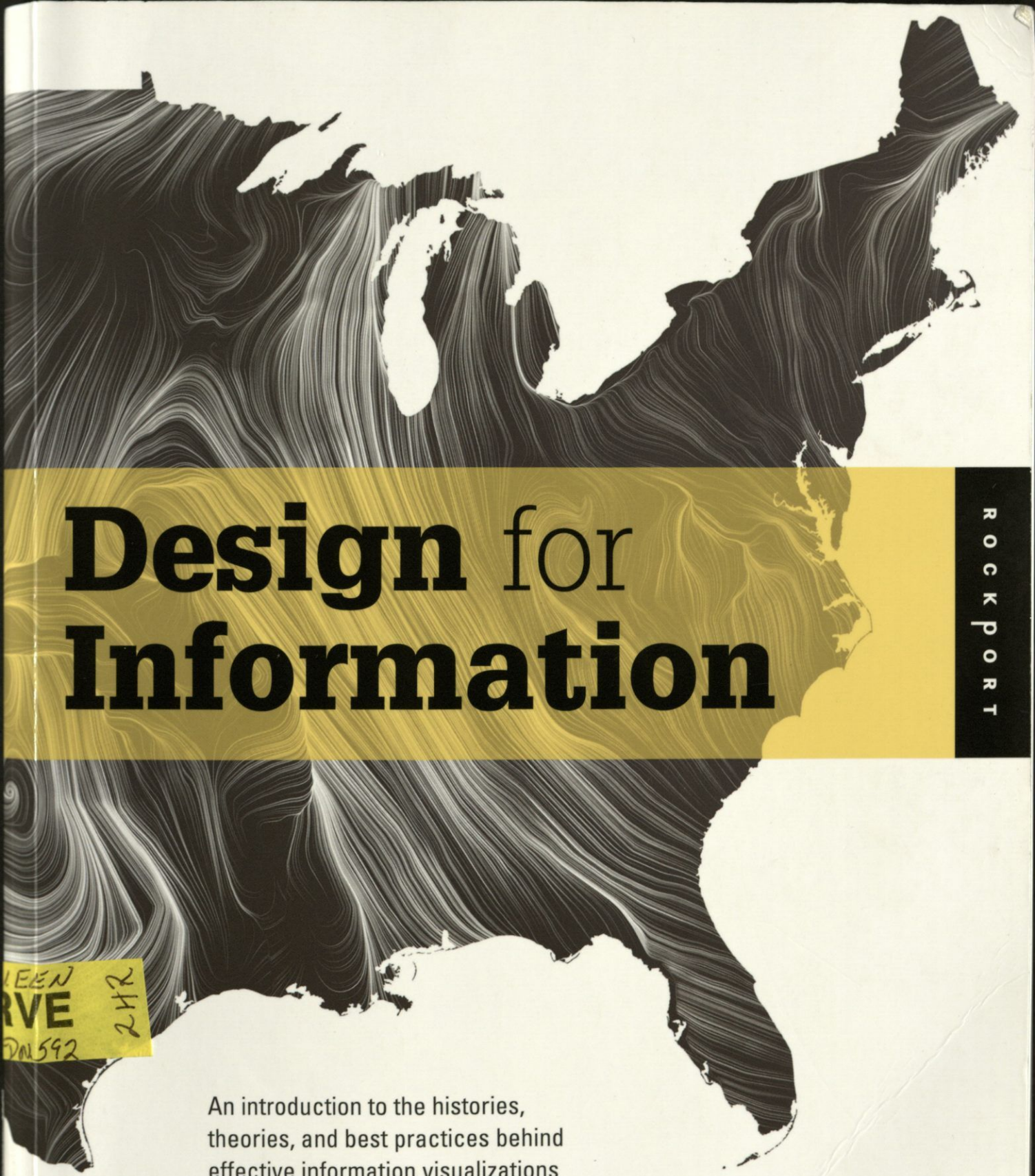


Design for Information

ROCKPORT

An introduction to the histories, theories, and best practices behind effective information visualizations

Isabel Meirelles

First published in the United States of America in 2013 by
Rockport Publishers, a member of
Quarto Publishing Group USA Inc.
100 Cummings Center
Suite 406-L
Beverly, Massachusetts 01915-6101
Telephone: (978) 282-9590
Fax: (978) 283-2742
www.rockpub.com
Visit RockPaperInk.com to share your opinions, creations,
and passion for design.

10 9 8 7 6 5 4 3

ISBN: 978-1-59253-806-5

Digital edition published in 2013
eISBN: 978-1-61058-948-2

Library of Congress Cataloging-in-Publication Data available

Design: Isabel Meirelles
Cover Image: "Wind Map" by Fernanda Viégas and
Martin Wattenberg

Printed in China

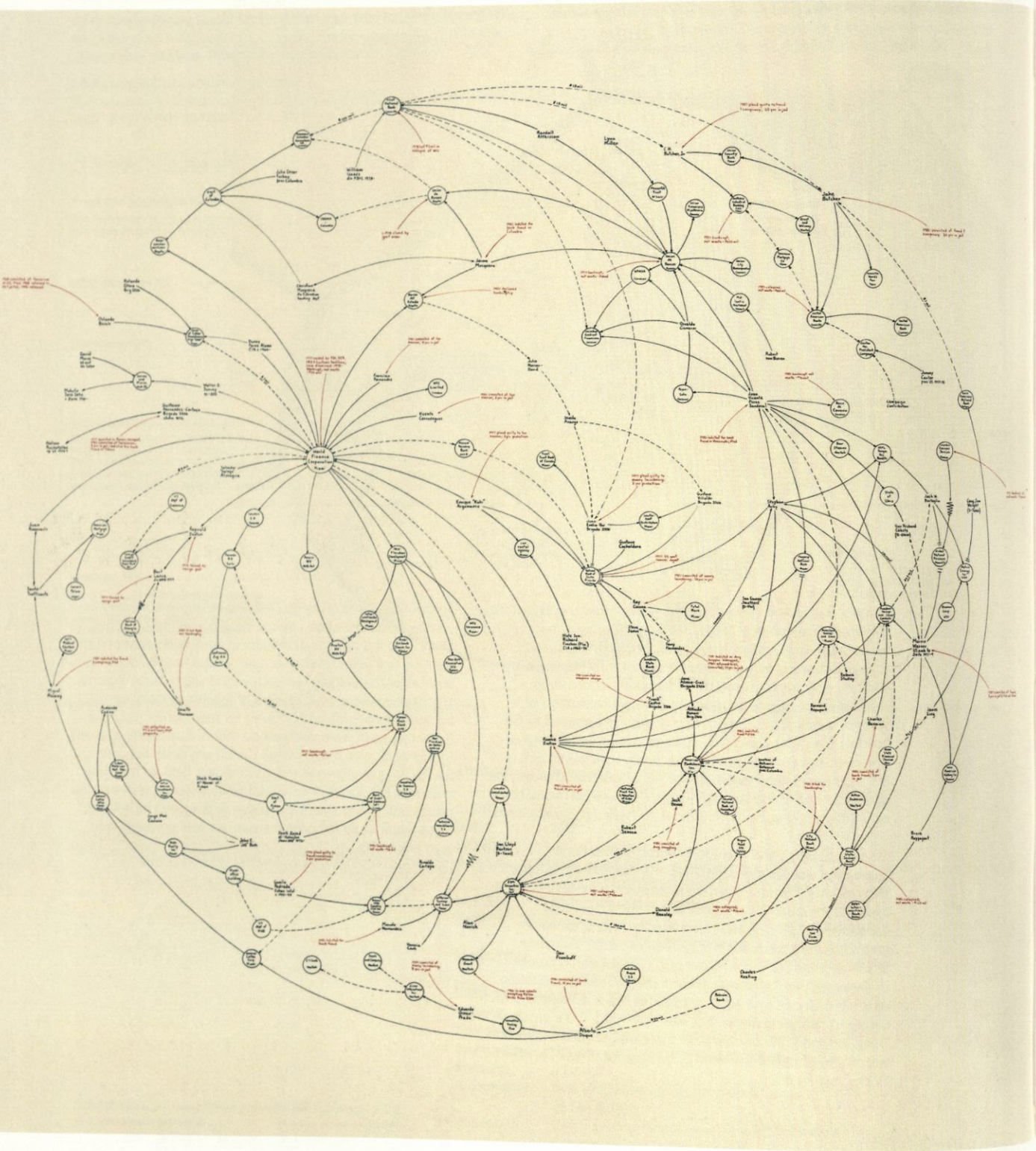
Design for Information

An introduction to the histories, theories, and best practices behind effective information visualizations

Isabel Meirelles

Rockport Publishers
100 Cummings Center, Suite 406L
Beverly, MA 01915

rockpub.com • rockpaperink.com



CHAPTER 2

RELATIONAL STRUCTURES: NETWORKS

Mark Lombardi, U.S.: *World Finance Corporation and Associates, c. 1970–84, Miami-Ajman-Bogota-Caracas* (7th version), 1999.

American artist Mark Lombardi began a series of drawings in 1994 that he called Narrative Structures, as he explains, "Each consists of a network of lines and notations which are meant to convey a story, typically about a recent event of interest to me, like the collapse of a large international bank, trading company or investment house. One of my goals is to map the interaction of political, social and economic forces in contemporary affairs."[13] The drawing reproduced here is one among several versions Lombardi created to depict the scandal involving the WFC and the central role it reputedly played in the trafficking of Colombian drugs. Robert Hobbs explains, "An important subtext of this work and other Lombardi pieces...is the wide-ranging collusion involved in global crimes."[14]

As the name indicates, relational structures organize data for which relationships are key to the system being visualized. Or to put it another way, there is much that can be learned by studying the patterns of connections between elements in the system—that is, the network of systems. Shneiderman and colleagues provide a good example in the context of social studies: "The focus of social network analysis is between, not within people. Whereas traditional social-science research methods such as surveys focus on individuals and their attributes (e.g., gender, age, income), network scientists focus on the connections that bind individuals together, not exclusively on their internal qualities or abilities. This change in focus from attribute data to relational data dramatically affects how data are collected, represented, and analyzed. Social network analysis complements methods that focus more narrowly on individuals, adding a critical dimension that captures the connective tissue of societies and other complex interdependencies."[1]

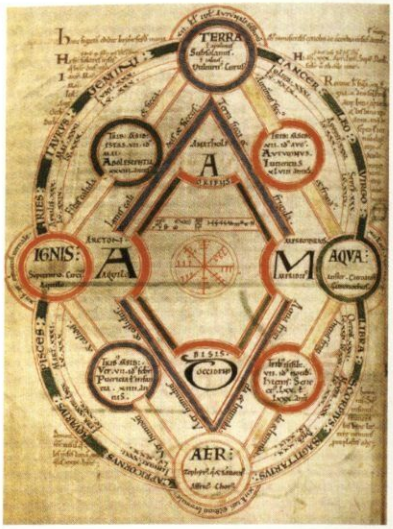
We are surrounded by networks, from metabolic to social networks, from transportation systems to power grids. Barabási explains that networks are at the heart of understanding complex systems, and "despite the apparent differences, the emergence and evolution of different networks is driven by a common set of fundamental laws and reproducible mechanism. Hence despite the amazing diversity in form, size, nature, age, and scope characterizing real networks, most networks observed in nature, society, and technology are driven by common organizing principles."[2]

The study of networks is not new, as shown by early research in fields as varied as biology, sociology, and mathematics, briefly described below. The scientific study of networks—network science—is, however, more recent and focuses on the study of patterns of connections in real-world systems. According to Barabási, four key characteristics distinguish network science as a discipline from early studies of networks: it is interdisciplinary; it examines empirical data; it is quantitative and mathematical in nature; and it relies on computational tools.[3]

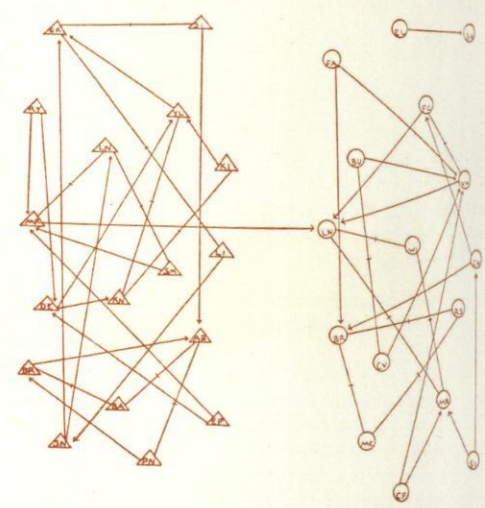
Over the years, scientists from several fields have developed a set of tools for analyzing, modeling, and making predictions about complex systems using network science. Given the mathematical, computational, and statistical nature of these tools, this book offers only a glimpse into the fundamentals of this burgeoning field, with focus on visualizations. Visualizations have played a key role in network sciences by adding visual insight and intuition to the numerical analysis. By examining network representations from diverse disciplines, I present the core concepts of network analysis while discussing the challenges faced in visualizing them. For an in-depth examination and advanced study of the science behind networks, I recommend excellent books listed in the bibliography (see page 209). Important to remember is that new models and algorithms are constantly being devised by a growing number of researchers all over the world, and those can be found in scholarly papers and conference proceedings.

GRAPH THEORY

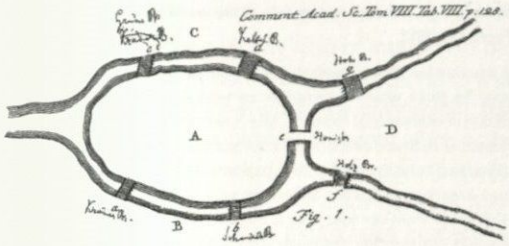
Network science originated in graph theory, and the mathematical foundations set by Leonhard Euler in the eighteenth century. The root is on the puzzle involving the city of Königsberg, the capital of eastern Prussia at the time, and its seven bridges: Can one walk across all seven bridges without crossing the same bridge twice?

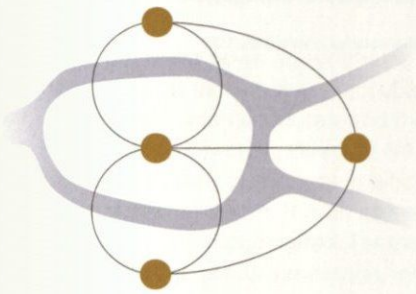


This diagram by Byrthferth de Ramsey, from a manuscript from around 1080, portrays the mysteries of the universe. It shows Adam in the center, surrounded by the four cardinal points (north to the left), the four elements, the four seasons, the four stages of life, and the twelve signs of the zodiac.



Considered one of the founders of social network analysis, the Romanian psychiatrist Jacob Moreno devised a method for evaluating relationships between individuals in groups or communities called *sociometry*. *Sociograms* are the visual counterpart he devised to represent information as graphs for studying the connecting roles of individuals in communities. His 1934 book *Who Shall Survive?* presents his theories and early network graphs.





The image at the top appeared in the original paper by Euler in 1736, and it shows the seven bridges in the city of Königsberg. The diagram at the bottom depicts the same problem as a graph.

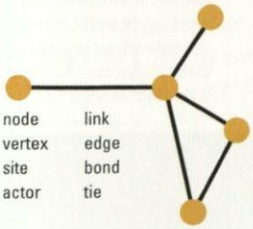
In 1736, Euler provided a mathematical proof showing that the path didn't exist. Euler's proof was the first time someone solved a mathematical problem by turning it into a graph, where land areas were represented as nodes and the bridges as links. Euler observed that except for the starting and ending nodes, all other nodes should have two links—in and out—or an even number of links, if an *Eulerian path*, as it came to be called, is to exist. Described in another way, "a network can have an Eulerian path only if there are exactly two or zero vertices of odd degree—zero in the case where the path starts and ends at the same vertex."[4]

BASIC ELEMENTS

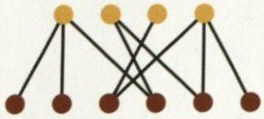
Networks are collections of nodes and links with a particular structure, or topology. A network is also called a *graph* in mathematics. As Newman explains, "A network is a simplified representation that reduces a system to an abstract structure capturing only the basics of connection patterns and little else. Vertices and edges in a network can be labeled with additional information, such as names or strengths, to capture more details of the system, but even a lot of information is usually lost in the process of reducing a full system to a network representation."[5]

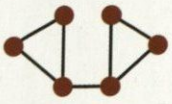
A node can be a machine, a person, a cell, and so on. A link represents the relations between two nodes. For example, in the Internet, nodes are computers or routers, and links are cables or wireless connections; in a neural network, nodes are neurons and links are the synapses. Different disciplines use different terminology to describe the elements of networks, but differences stop at the label conventions.

FIELD		
Computer Science	node	link
Mathematics	vertex	edge
Physics	site	bond
Sociology	actor	tie

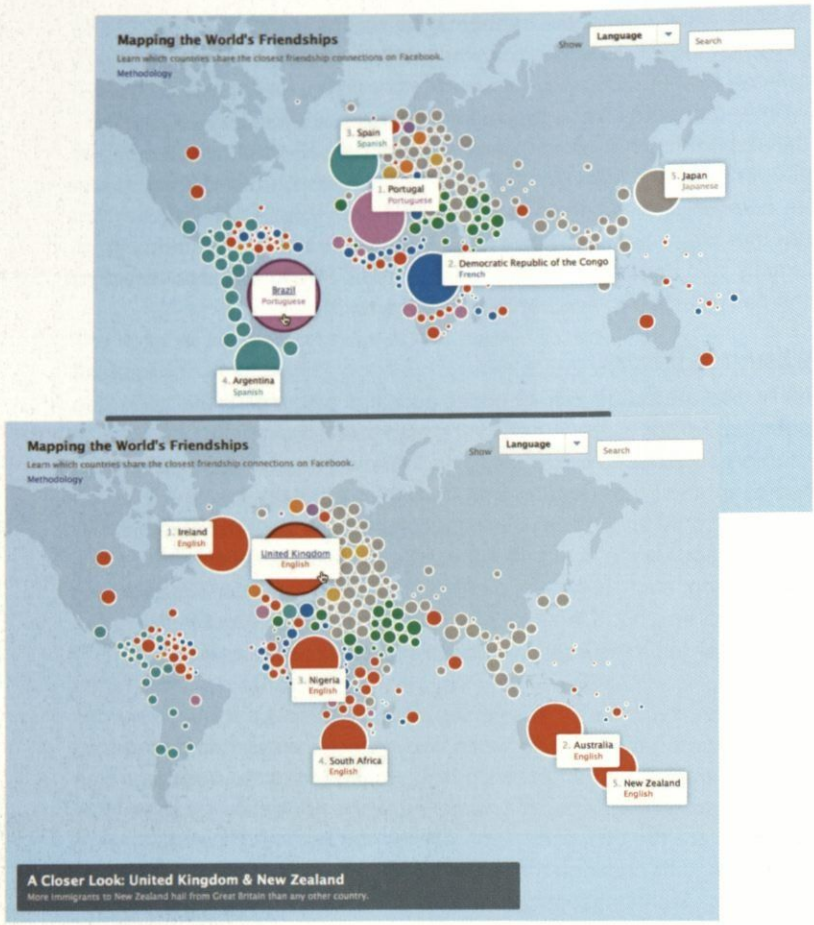


When all the nodes in the network are of the same type, say, in a friendship network, where all nodes are friends, the network is called *one mode*, *unimodal*, or *unipartite*. When there is more than one type of node, networks are called *multimodal* or *multipartite*. A common examined type is the bipartite network, also called a *two-mode* or an *affiliation network* (in sociology), which consists of two sets of nodes that only share links between sets, but not within them. For example, a network with two sets of nodes, persons and books, and the links showing who has read what is considered a bipartite network. Out of this network, we can construct two one-mode projections: person–person, where a node is a person and the nodes are connected if they have read the same books; and book–book, in which books are connected if they share the same readers. One-mode projections allow understanding of clusters based on common membership.





The center graph shows a bipartite network, and the top and bottom ones are the related one-mode projection.



Stamen Design, U.S.: Interactive map of the world's friendship in Facebook, 2012.

The visualization shows trends in how friendship ties are dispersed around the globe, which were further contextualized by Stanford graduate in international relations, Mia Newman. The top image shows connections to Brazil, and it reveals, for example, a strong relationship with Japan. This is due to migration patterns, and more specifically, a large Japanese emigration to Brazil more than 100 years ago. The bottom image shows connections to the United Kingdom, this time with colors representing languages. This image shows strong ties with countries that once were Britain's colonies. A similar pattern is also found in connections of former empires, such as in Portugal and France.

www.facebookstories.com/stories/1574

Paul Butler (Facebook), U.S.: Map visualizing friendships in Facebook around the globe, 2010.

Perhaps what is most revealing about the image are the dark spots—in other words, the lack of connections in certain areas. It turns out, however, that those places are not uninhabited, nor isolated technologically; rather, these places are inhabited by populations with different choices of social media applications.

PROPERTIES

Links are described by any kind of interaction between nodes, from kinship to collaboration, from transactions to shared attributes. For example, in a social network, the links might stand for different kinds of interactions between people, such as family, friend, work-related, political affiliation, etc. In a trade network, with countries as the nodes, links might stand for types of transactions, such as import or export. Links might have properties describing the direction of the interaction (undirected or directed), and the weight of that connection (unweighted or weighted).

Undirected links, also known as symmetric edges, refer to mutual connections, such as those between couples. Undirected links have no origin destination attributes, and the lines are represented without indication of direction. Directed links, also known as asymmetric edges, are connections in which an origin destination between the nodes is known. It is an asymmetric relation because not all connections are reciprocated. For example, when someone makes a phone call or sends a message, we can identify who originated it (*from*), who it was designated for (*to*), and whether it was or was not reciprocated. In ecological networks, such as food webs, directed links show the prey-predator interactions. Directions are often represented with the addition of arrows to the link elements.

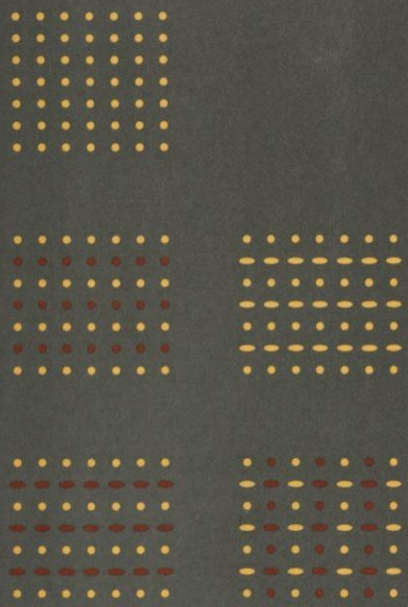
Unweighted links describe the existence of a connection without further indication of its nature. In other words, it is an on/off situation, where the presence of a link between two nodes denotes that there is an interaction between them without other qualifications. Weighted links, on the other hand, represent additional information about the interaction, such as its strength, weight, or value. For example, John calls Mary more frequently than he calls Joseph. The weight is often represented by the quality of the line, and most often quantities are represented by its width.

The number of immediate connections of a node provides the degree property of that node. In the case of directed networks, degrees are designated as "in degree"—the number of links destined *to* the node, and "out degree"—the number of connections originated at the node, or *from* it. Once we know the degree of a node, there are other metrics that can be analyzed, such as the notion of the degree centrality of a node in relation to the network, which attempts to answer questions about the "importance" of that node in the system. There are few ways that centrality can be measured, from the simple calculation of the node degree in

Similarity

Similarity is the tendency to group similar visual elements into a perceptual unit. It relates to nonlocational characteristics, such as color, shape, and texture, but it is not absolute and can occur in degrees. For example, we perceive the six lines below as forming different groups:

|||||| = 1 group
|||||_ = 2 groups



Similarity is essential to categorical association. For example, the use of color coding for categories can enhance the search and comparison between them. When associating more than one graphical variable, attention should be paid to whether they are integral or separable visual dimensions.

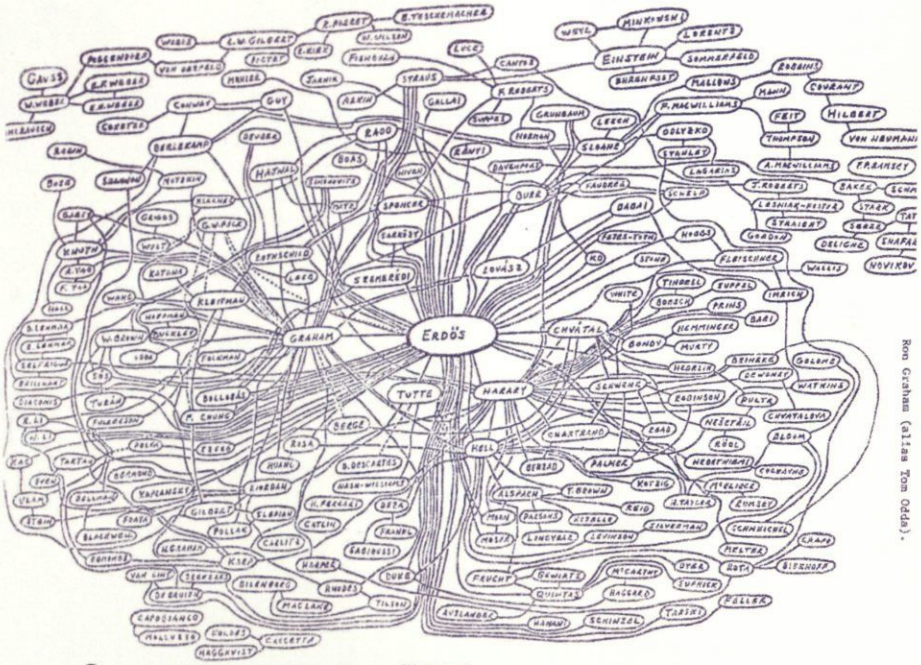


Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

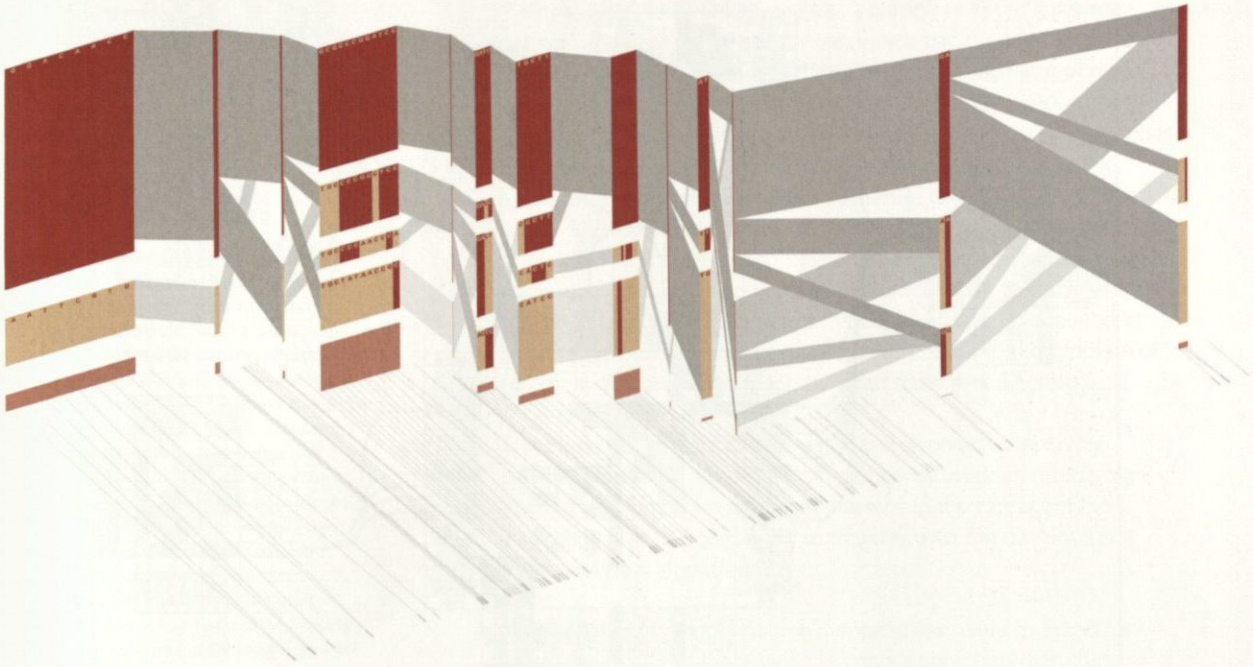
Paul Erdös published around 1,500 papers during his prolific career in mathematics, including important contributions to graph theory and random graphs. Ron Graham hand drew this diagram in the 1970s to portray the collaboration network of Erdös. The nodes are mathematicians, and links connect pairs who have jointly authored a paper with Erdös. As Easley and Kleinberg explain, "A mathematician's Erdös number is the distance from him or her to Erdös in this graph. The point is that most mathematicians have Erdös numbers of at most 4 or 5, and—extending the collaboration graph to include co-authorship across all the sciences—most scientists in other fields have Erdös numbers that are comparable or only slightly larger; Albert Einstein's is 2, Enrico Fermi's is 3, Noam Chomsky's and Linus Pauling's are each 4, Francis Crick's and James Watson's are 5 and 6, respectively. The world of science is truly a small one in this sense."[15]

relation to the total number of links in the network, to more complex methods involving paths, such as closeness and betweenness centrality. A common method is the eigenvector centrality, which considers the importance of nodes based not only on how many links the node has but also in relation to the degree centrality of its neighbors.

Another example is the study of degree distribution, which became a central measure after the discovery of the scale-free networks in 1999 by Albert and Barabási.[6] The model shows that the average number of connections follows a power law distribution: many nodes with few connections (small degree) and a few nodes with many connections (very high degree).

PATHS AND CONNECTIVITY

In order to understand distances in a network, scientists developed the concept of a path that is any sequence of nodes given that each consecutive pair of nodes is connected by a link. The path length provides the number of links in the route between a pair of nodes. When there are no paths between a pair of nodes, it means that a network is not connected and it is divided into subgroups, called "components" in network science. Other metrics were devised for questions related to distances; the shortest path between two nodes and the network diameter are two examples.



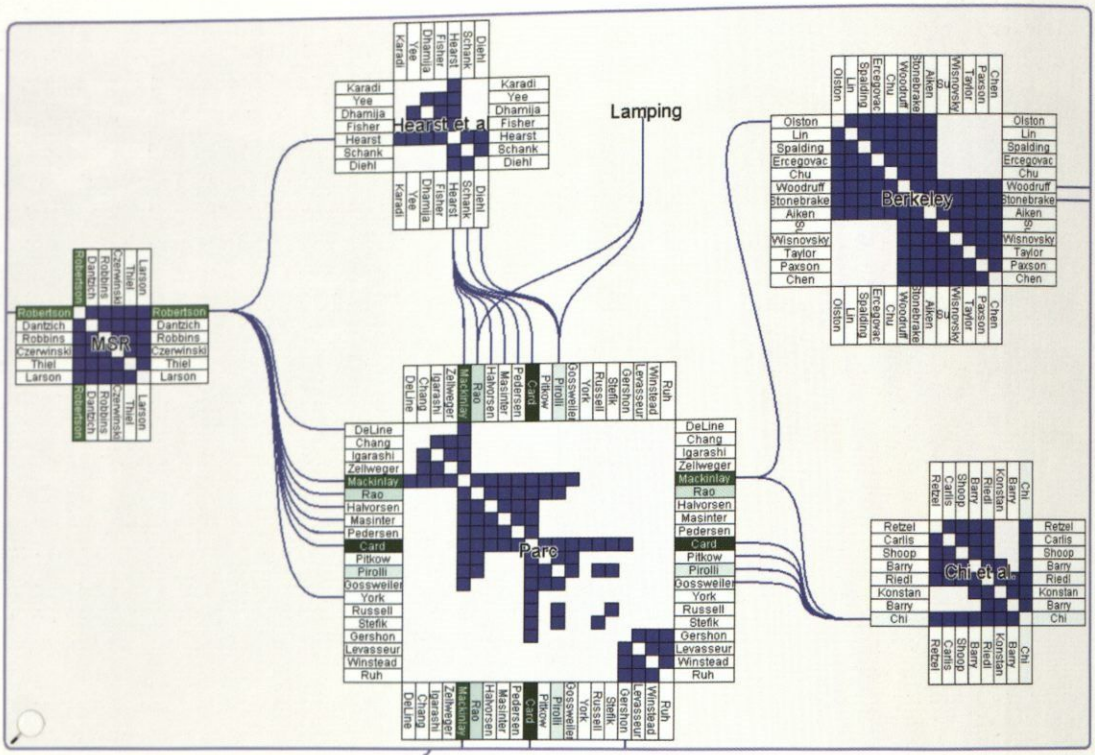
Ben Fry, U.S.: "*Isometricblocks,*" 2003.

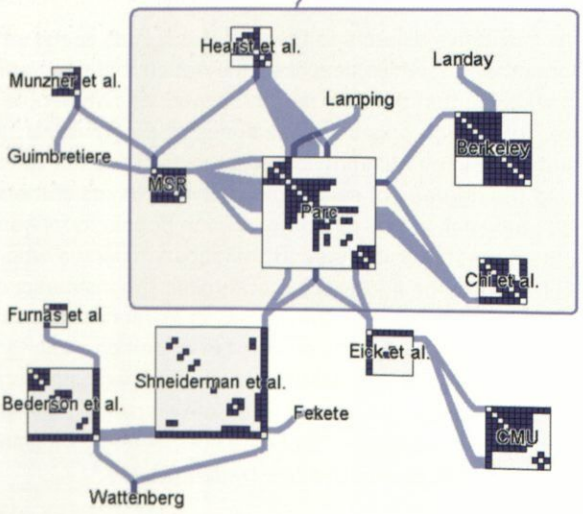
Ben Fry developed this interactive applet using the open source Processing environment, which he originally conceived and implemented with Casey Reas. His goal was to devise software that would combine scientific methods with visualization tools for haplotype and LD data. One can interactively switch between methods and look at the different visualization options. Animated transitions make comparisons easy to follow. One can also modify the parameters of the mathematics used to set boundaries on the blocks. Unfortunately, this caption is far from capturing the level of sophistication of this application, and further reading on how Fry developed it is strongly recommend by following the URL below. This screenshot shows the 3-D view option, an isometric projection, with blocks offset slightly in the *z*-axis to allow view of the lines depicting the transitions between blocks, while preserving the linear scaling of the nucleotide scale in the horizontal axis.

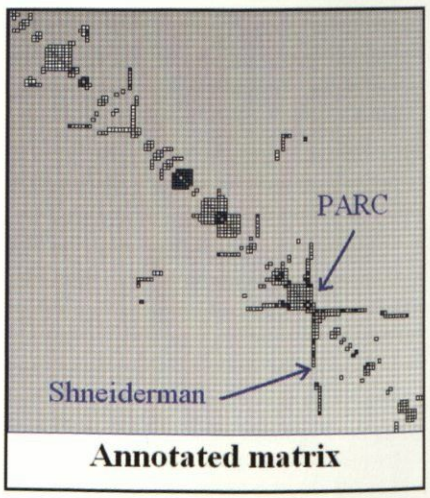
www.benfry.com/isometricblocks

There are qualitative aspects to these metrics, such as the small-world phenomenon, which describes the notion that the world feels small, given that the path length connecting two people is a short one. This notion originated in a series of experimental studies performed by Stanley Milgram and colleagues in the 1960s, which looked into the degrees of separation between chains of friends.[7] One such study asked 296 randomly chosen people to forward a letter in the fastest possible way to a destination person who lived in the suburbs of Boston and was a stockbroker. The instructions were to give the letter to someone they knew on a first-name basis and who would be more likely to know the target person. Among the 64 letters that arrived at the destination, the median path length was six. Later, this phenomenon came to be known by the popular notion of "six degrees of separation," a phrase not coined by Milgram but inspired by the 1990 play of this title by John Guare.[8]

Many network metrics have originated in the social sciences, and are now commonly utilized for quantifying network structures across many fields. The metrics and models have helped scientists examine networks in different domains while also looking for universal properties underlying these phenomena.







Annotated matrix

Nathalie Henry Riche, Howard Goodell, Niklas Elmqvist, and Jean-Daniel Fekete, France: *NodeTrix*, 2007.

Following the pioneering work with reorderable matrices of Jacques Bertin, Riche and colleagues devised several tools to explore and understand networks in the digital environment. *NodeTrix* was directed at solving the problem of how to represent networks that are globally sparse with dense local communities. The interactive tool uses a hybrid representation that combines matrices, for depiction of dense areas, within a node–link diagram, which provides the global structure.[16]

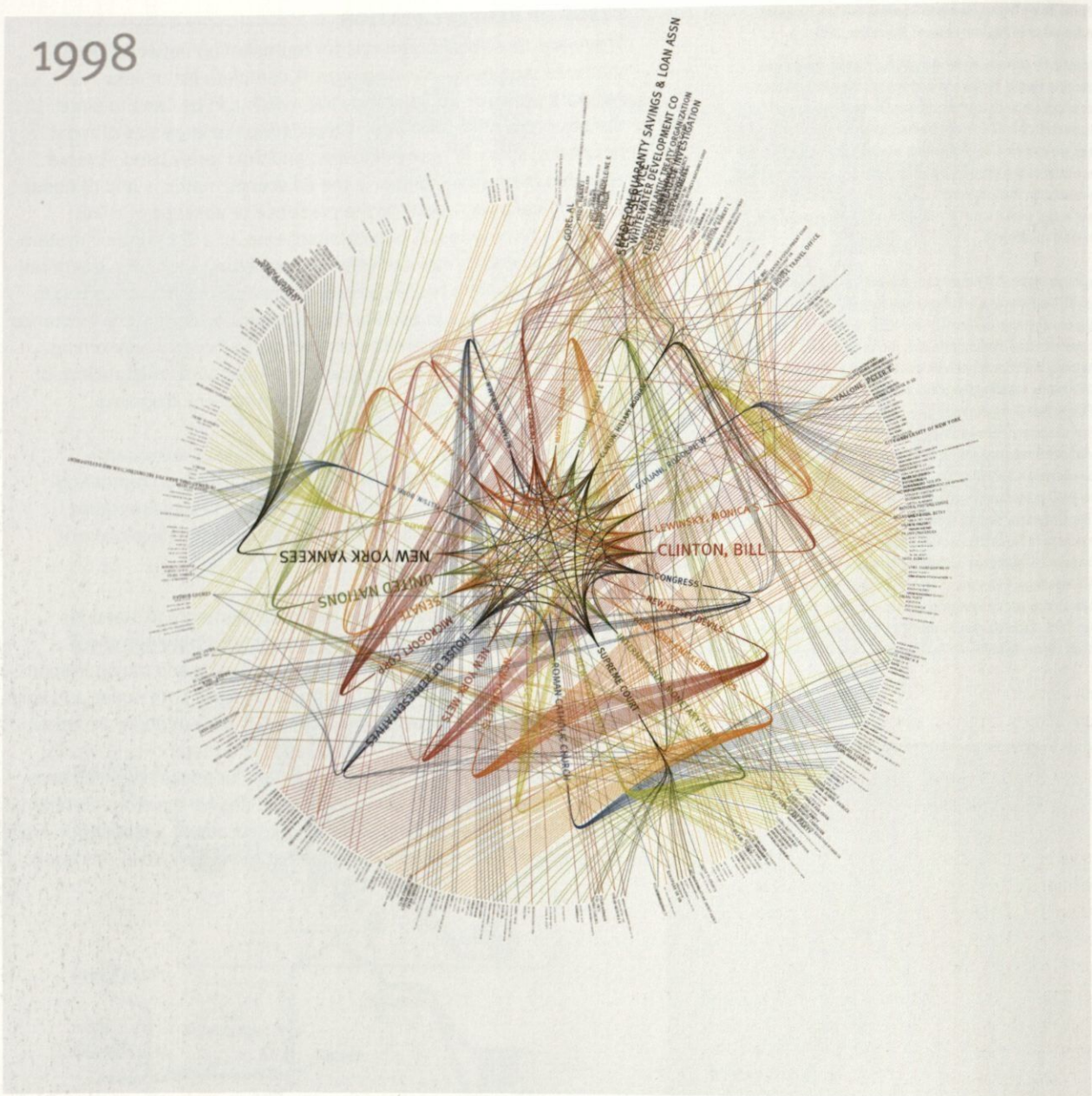
The image is part of a larger examination of "20 Years of Four HCI Conferences." It shows the largest component of the coauthorship network of the IEEE Symposium on Information Visualization (InfoVis). Riche explains, "The lower right corner shows the overview of whole InfoVis matrix, labeling the main actors of this network: PARC and Ben Shneiderman. The largest cross identifiable is Ben, the most central actor in the InfoVis community. The *NodeTrix* representation in the lower left corner shows how Ben Shneiderman acts as a bridge to the other UMD researchers grouped in a community centered on Ben Bederson. Finally, the upper part of the figure is a zoomed-in *NodeTrix* view showing how the PARC community collaborates with other communities. It is interesting to note that Berkeley and Microsoft Research strongly collaborate with each other. Similarly Stuart Card, Jock Mackinlay and Ed Chi collaborators are strongly connected."[17]

TYPES OF REPRESENTATION

There are three main methods for representing networks: lists, matrices, and node–link diagrams. A complete list of links in a network provides an adjacency list, which can be used to store the structure of the network. Considering the large size of most networks, lists are unmanageable, and thus rarely used. A more effective mathematical tool is the adjacency matrix, a grid of nodes with the cells representing the presence or absence of a link between two nodes. A two-color scheme, or a 0/1 numeric system, usually suffices to indicate links in unweighted networks. Weighted networks require a more complex numerical or visual encoding to represent amounts in addition to the binary system of the existence of a link. One of the benefits of matrices is that by representing information about links in the cells, matrices avoid the problem of too many link crossings faced by most node–link diagrams.

The French cartographer Jacques Bertin worked extensively on matrices in the 1960s.[9] Bertin pioneered work on reordering rows and columns for revealing patterns in the representation, a strategy that has continued to this day with the development of several new algorithms.

Node–link representations use symbolic elements to stand for nodes, and lines to represent the connections between them. Physical network systems, such as power grids and transportation networks, provide the spatial attribute to locate both nodes and links into the spatial structure of the diagram. Most networks, however, are of abstract data, such as food webs and metabolic networks, and do not have *a priori* spatial properties for positioning elements in the visualization. The table on page 62 shows the most common types of layouts according to certain properties of the network. Each type points to real-world examples that are examined in this book.



Jer Thorp, U.S.: *"New York Times 365/360,"* 2009.

The image is one of a series of visualizations that Canadian Jer Thorp created in Processing for portraying the top organizations and personalities for every year from 1985 to 2001, by occurrence in the *New York Times*. Lines indicate the connections between people and organizations.

CHALLENGES OF NODE–LINK DIAGRAMS

Most problems faced by node–link representations are caused by the occlusion of nodes and link crossings, which obliterates the structure it is supposed to reveal. This, however, is not a trivial problem, given the large datasets used in these graphs, and, perhaps, one of the reasons we so often see hairball network displays, which are hard to read and extract meaning from. New layout techniques and algorithms aim at minimizing the problem, and ultimately generating more legible graphs. This is an area receiving large attention by researchers in diverse fields, because the need for effective visual displays of networks grows with the accessibility to more and larger datasets. As Schneider contends, "It could be said that a graph is worth a thousand ties.... The ability to map attribute data and network metric scores to visual properties of the vertices and edges makes them particularly powerful. However, network visualizations are often as frustrating as they are appealing. Network graphs can rapidly get too dense and large to make out any meaningful patterns. Many obstacles like vertex occlusions and edge crossings make creating well-organized and readable network graphs challenging."[10]

One of the strategies pursued in interactive node–link visualizations is the ability to switch between different spatial layouts in order to discover meaningful properties of the network, while understanding relationships in new ways. Take, for example, the *SPaTo* application tool that allows examination of properties of a network by switching from a force directed node–link representation, to a circular graph, to a geographical map (see pages 76–77).[11] Similar to reordering rows and columns in matrices, the rearrangement of the spatial layout helps revealing hidden structures in the network.

Visual encoding of nodes and links is another area that affects the interpretation of network representations, in that complex systems are often described by more properties than we can perceive. Take, for example, the Human Disease Network, which is an amalgamation of more than seven subsystems—our limit to perceptually distinguish and cognitively remember stuff (see the box Magical Number Seven on page 97). However, eliminating categories from the eighteen disorder classes portrayed in the visualization would negatively affect the integrity of the information being communicated: "A platform to explore in a single graph theoretic framework all known phenotype and disease gene associations, indicating the common genetic origin of many diseases."[12]

What about Trees?

Newman describes a tree as "a connected, undirected network that contains no closed loops.... A network can also consist of two or more parts, disconnected from one another, and if an individual part has no loops it is also called a tree. If all the parts of the network are trees, the complete network is called a *forest*."[18] Nodes that are connected to other nodes are called *leaves*.

A tree is considered a connected network because every node can access any other node by following a path. Given that there are no loops in trees, there is only one possible path between any pair of nodes.

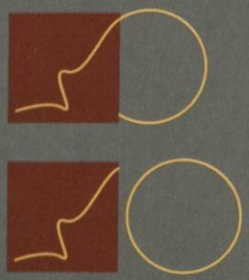
Topologically, a tree can be represented with any node as its root. There might be some specific reason for choosing a node as its root, such as what we saw in the previous chapter on hierarchical structures.

Examples drawn after Newman (2007), 127.

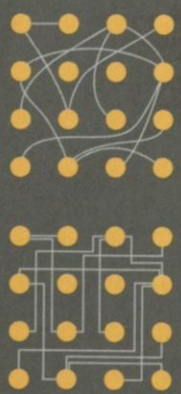
Good Continuation

Good continuation is the tendency to construct visual entities out of visual elements that are smooth and continuous, or connected by straight or smoothly curving lines. For example, we perceive the six lines below as forming different groups:

= 1 group
= 2 groups



A common experience of the principle is found in most maps. Good continuation allows, for example, for state contours to be differentiated from roads or rivers. When representing data, we should pay attention to not creating nonintentional groupings due to good continuation.

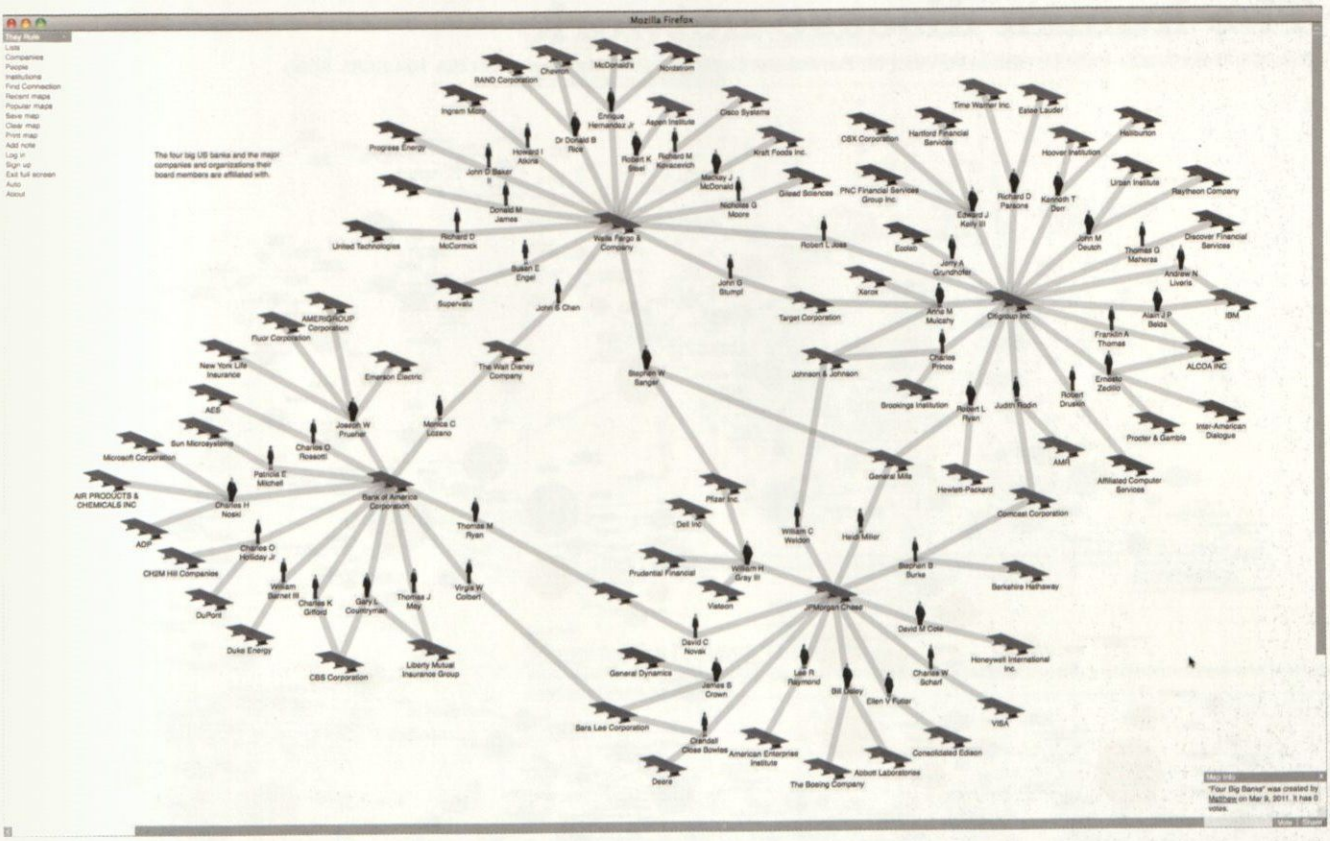


Good continuation plays an important role when designing networks with several connecting lines. It is easier to perceive smooth continuous lines than lines with abrupt changes in direction.

The same holds true for the placement of labels and the difficulties encountered in effectively positioning them, given the complexities of network representations. As with all other types of visualizations, labels carry important information, enabling one to understand what it is being revealed, from scales and measurements to categorical information. A common strategy in the case of node–link diagrams is placing labels inside the nodes. However, this is not always possible, such as in dense areas of the graph or in the case of small nodes with long labels. These limitations might be overcome in interactive visualizations, such as associating the cursor with actions that highlight nodes while revealing labels.

Other effective strategies involve enabling the user to change the camera view or zoom into the graph, for example. So-called *focus + context* techniques involve operations that keep the contextual view of the whole graph while enabling a selected area to be represented in detail. Presenting details as one gets closer is a strategy that has been used successfully in maps, in which the amount of details change in relation to the scale: the larger the scale, the greater the details, as in a neighborhood map, for example (see page 123).

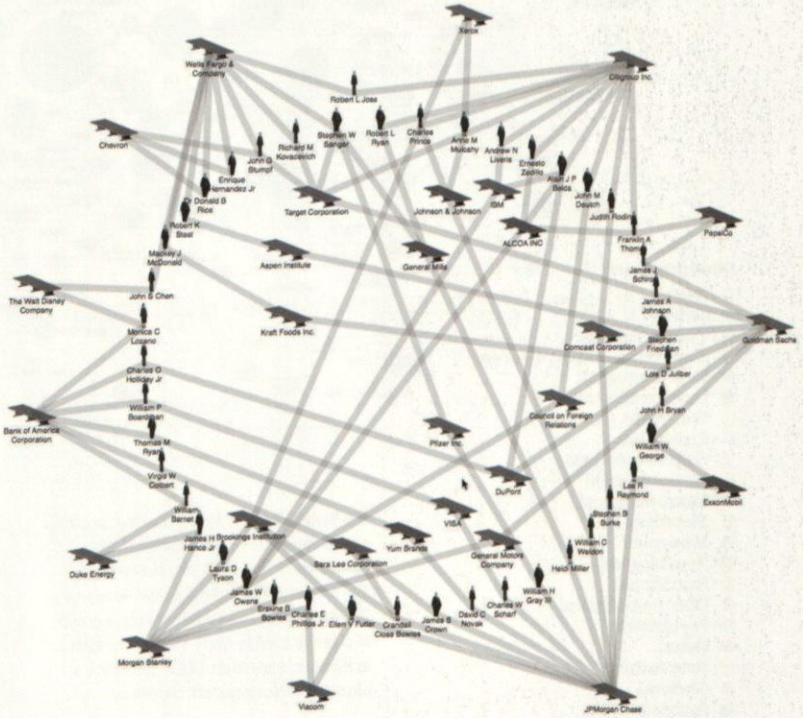
There are several mechanisms for reducing the number of nodes and links, such as using thresholds in the process of generating the visualization, collapsing nodes into clusters, or enabling one to filter data, three commonly used operations. The interactive network visualization in the website theyrule.net uses collapsing nodes, which can be revealed on demand by the viewer by selecting the group symbol (a table). In the series of images created by Thorp depicting data from the *New York Times*, links were bundled to avoid too many edge crossings in the circular layout.



Josh On, U.S.: "They Rule," 2004.

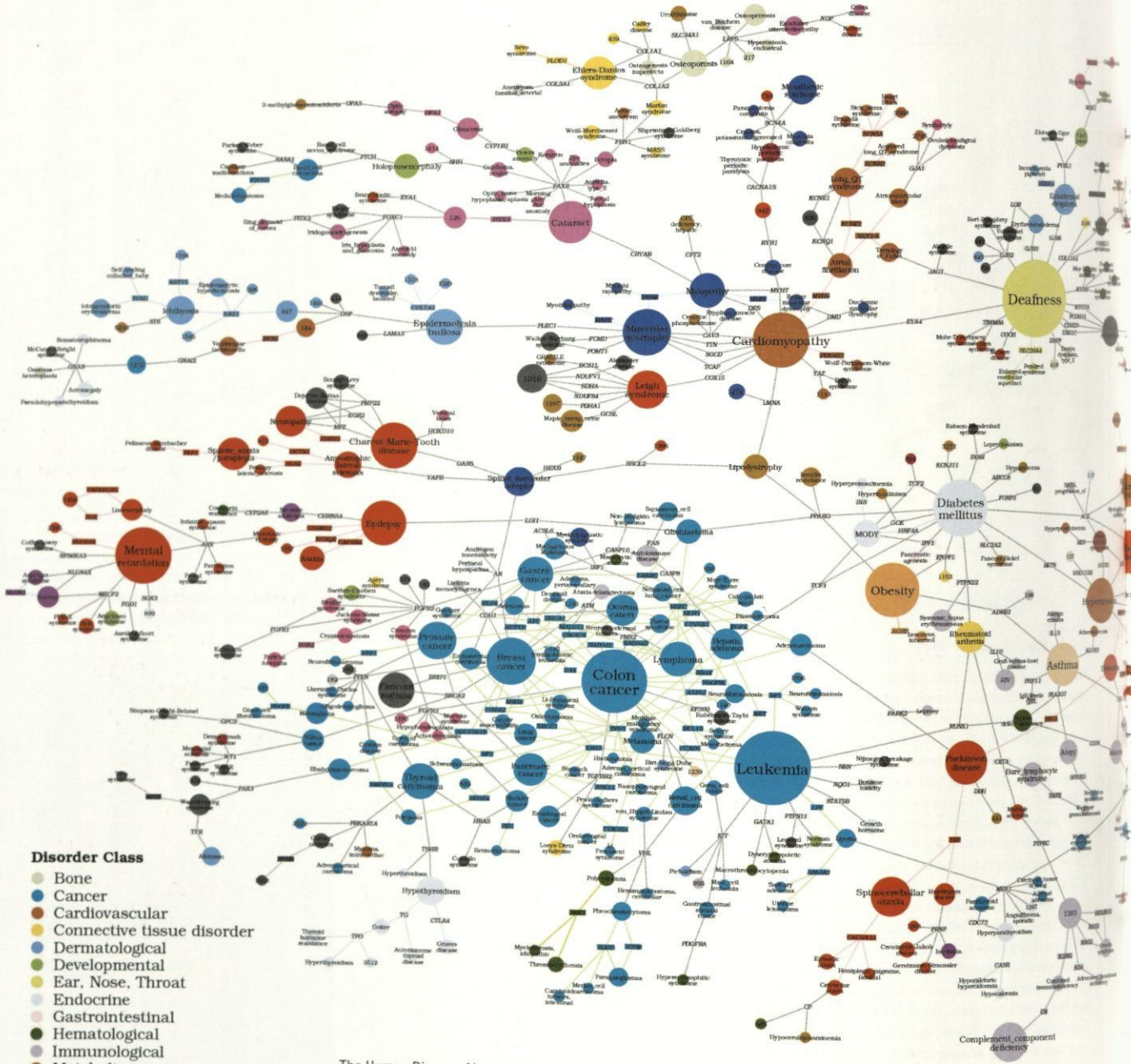
The website theyrule.net allows visitors to examine the connections among board members of the top 1,000 U.S. companies. It presents information in a relational diagram. Originally created in 2001 by Josh On with a static set of 100 companies, the site was updated in 2004 to include the top 500 companies, and in 2011, with data made available through LittleSis.org, the site now offers access to 1,000 companies. There are two types of nodes: organizations (table) and board members (male and female figures). Board members get "fatter" according to the number of boards they participate in. Corporation symbols do not change size, but they can be collapsed so as to hide board members in two ways: hide unconnected members or hide all members. Links connect board members to the organizations they serve on as well as among members when they sit on the same boards. Visitors to the site can save and share resulting graphs together with their own annotations. On writes about the context for building the tool: "Hopefully They Rule will raise larger questions about the structure of our society and in whose benefit it is run."[19] The images reproduced here were listed in the Popular Maps section. The one above is titled "Four Big Banks," and it was created by user Matthew on March 9, 2011, and the one on the right is titled "Six Too Big to Fail Banks," also by Matthew on July 26, 2011.

www.theyrule.net



The human disease network

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) *Proc Natl Acad Sci USA* 104:8685-8690

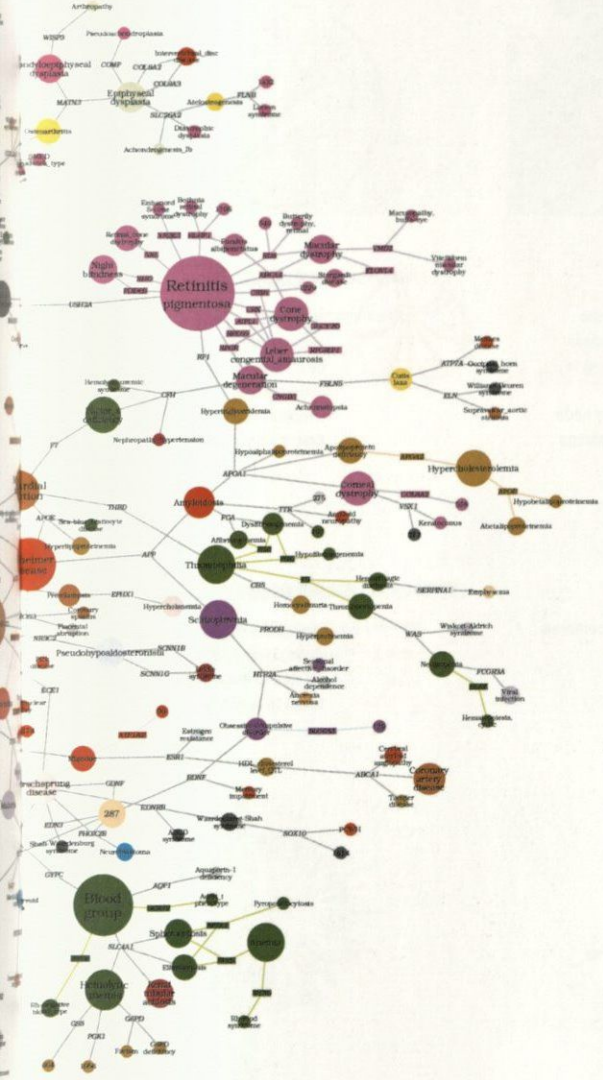


Disorder Class

- Bone
- Cancer
- Cardiovascular
- Connective tissue disorder
- Dermatological
- Developmental
- Ear, Nose, Throat
- Endocrine
- Gastrointestinal
- Hematological
- Immunological
- Metabolic
- Muscular
- Neurological
- Nutritional
- Ophthamological
- Psychiatric
- Renal
- Respiratory
- Skeletal
- multiple
- Unclassified

The Human Disease Network poster presents a bipartite-graph representation of the diseasome. It was originally published in *The Proceedings of the National Academy of Sciences* in 2007 and developed by a group of scientists led by Marc Vidal, a biologist at Harvard, and Albert-László Barabási, a physicist at Northeastern University.

A disorder (circle) and a gene (rectangle) are connected if the gene is implicated in the disorder. The size of the circle represents the number of distinct genes associated with the disorder. Isolated disorders (disorders having no links to other disorders) are not shown. Also, only genes connecting disorders are shown.[20]

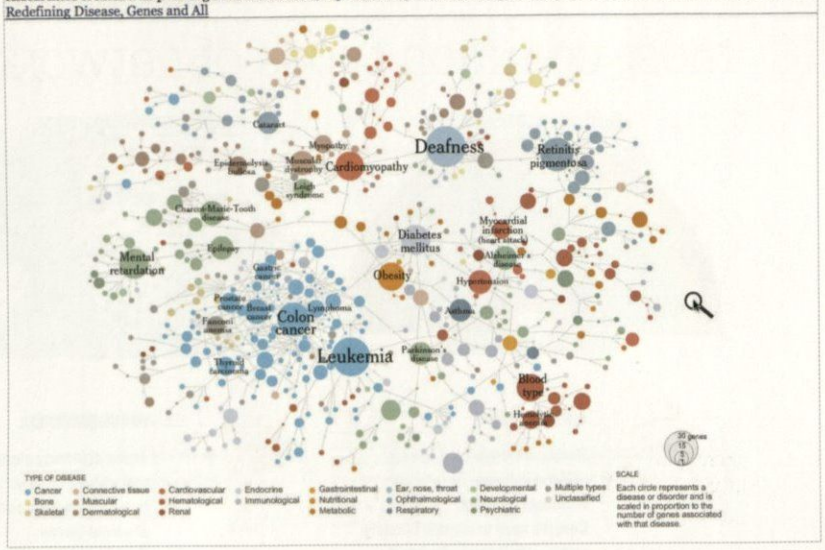


May 5, 2008

Mapping the Human 'Diseasome'

Researchers created a map linking different diseases, represented by circles, to the genes they have in common, represented by squares.

Redefining Disease, Genes and All

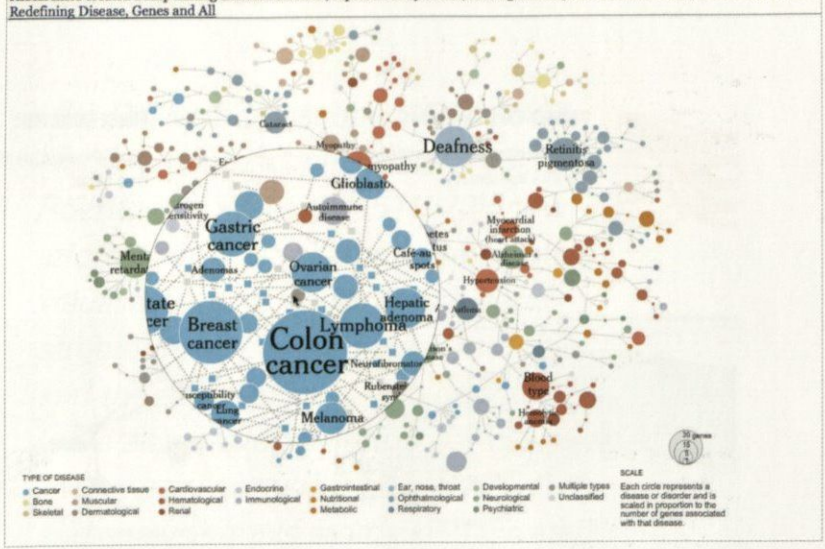


May 5, 2008

Mapping the Human 'Diseasome'

Researchers created a map linking different diseases, represented by circles, to the genes they have in common, represented by squares.

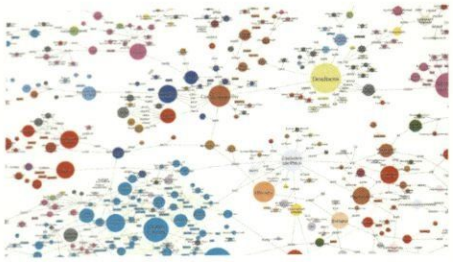
Redefining Disease, Genes and All

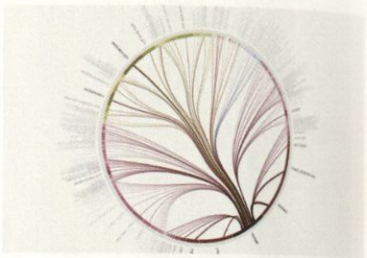


In 2008, the *New York Times* published the scientific discovery and created an interactive disease map to accompany the article "Redefining Disease, Genes and All."

www.nytimes.com/interactive/2008/05/05/science/20080506_DISEASE.html.

most common types of network layouts





LINEAR:

Nodes are organized linearly and the links are usually arcs connecting nodes.
Con: It's hard to identify clusters and is only feasible for small datasets.

FORCE DIRECTED:

There are many algorithms that use an iterative process to locate nodes according to physical forces.
Con: There are too many node occlusions and link crossings in dense areas.

CIRCULAR:

Nodes are organized around the circumference and usually grouped by categories. Links cross the circle and are usually bundled so as to simplify the crossings.
Con: It's hard to identify clusters.

SANKEY TYPE DIAGRAMS:

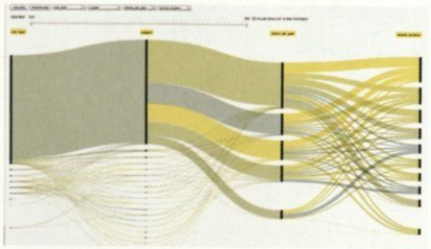
Nodes are organized vertically and the links horizontally.

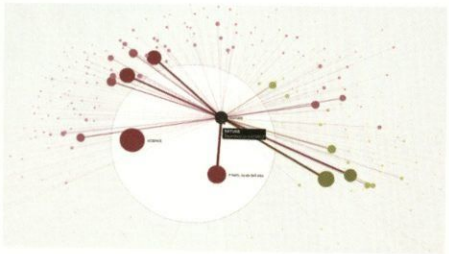
FORCE DIRECTED:

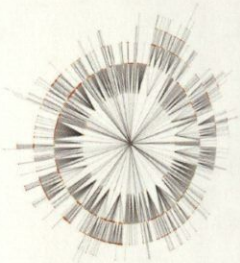
Force directed graphs centered on a node.

POLAR OR RADIAL:

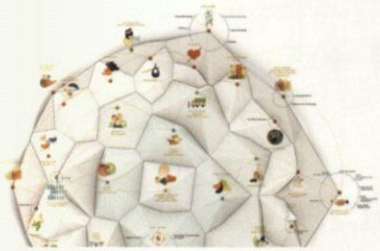
Nodes are organized around a central node, with their position related to the number of hops it takes to reach it.

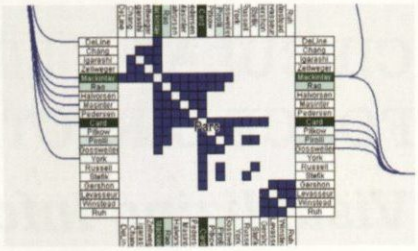






most common layouts centered on nodes



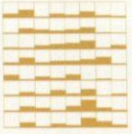


COMMUNITY STRUCTURE:

The focus is on community structures.

GEOGRAPHY BASED:

Spatial location of a node is provided by its geo position.



MATRIX:

Grid of nodes with link information positioned within the cell.

RADIAL COMMUNITY STRUCTURE:

Nodes are organized around a central community

Like Galileo's telescope (1564–1642), Hooke's microscope (1635–1703), or Roentgen's x-rays (1845–1923), new information analysis tools are creating visualizations of never before seen structures. Jupiter's moon, plant cells, and the skeletons of living creatures were all revealed by previous technologies. Today, new network science concepts and analysis tools are making isolated groups, influential participants, and community structures visible in ways never before possible.

Ben Shneiderman

CIRCULAR + LINEAR + TREEMAP + FORCE DIRECTED

Visualizing Information Flow in Science

http://well-formed.eigenfactor.org

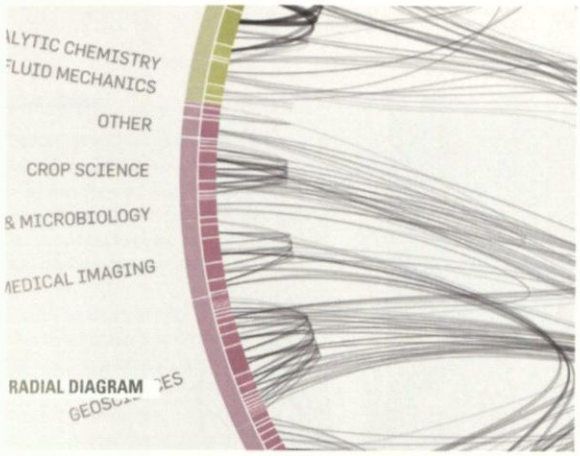
The visualization "well-formed.eigenfactor: Visualizing information flow in science" was devised by Moritz Stefaner in collaboration with Martin Rosvall, Jevin West, and Carl Bergstrom at the Bergstrom Lab, University of Washington.

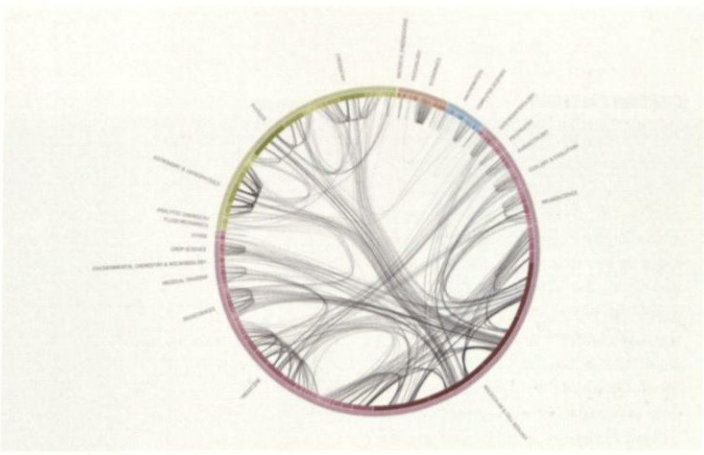
The project examines a subset of the citation data from Thomson Reuters' *Journal Citation Reports* from 1997–2005. It depicts 400 journals and around 13,000 citation edges, which ensures coverage of the top journals in each field.

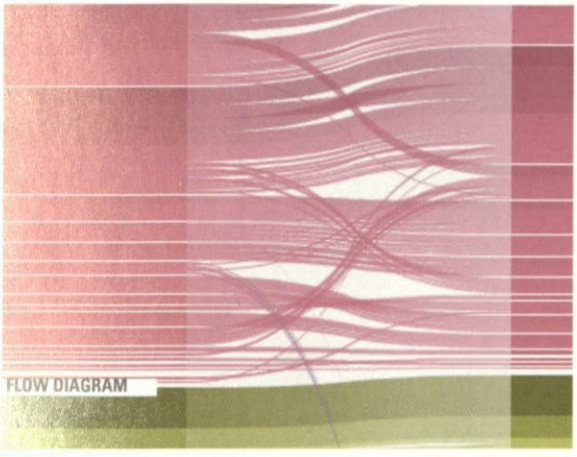
The Eigenfactor® project calculates a measure of importance for individual journals—the Eigenfactor score—while also measuring the citation flow and a hierarchical clustering based thereon. The authors explain how they approached the problem of visualizing the citation network, saying, "Our project extends the visual vocabulary in this area: on the one hand, by repurposing existing techniques, such as radial edge bundling and treemaps; on the other hand, by inventing novel approaches like magnetic pins as flow indicators and an alluvial diagram to represent change over time in cluster structure."[21]

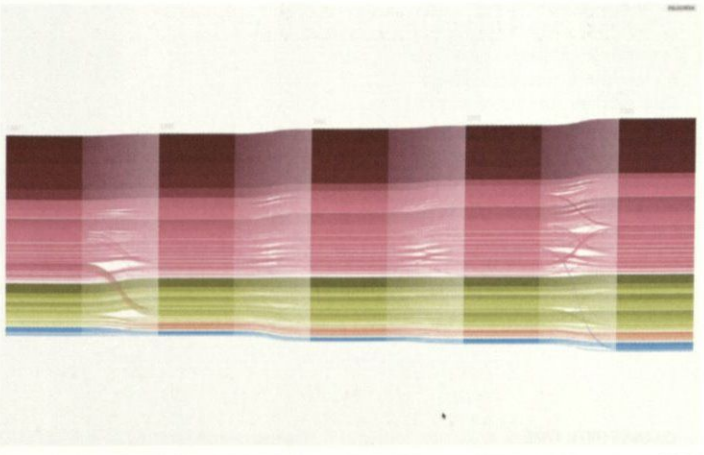
Stefaner created a set of four interactive visualizations, and each allows one to explore emerging patterns in the scientific citation network: citation patterns, changes over time, clustering, and map. All visualizations were created in 2009 using Flare, the ActionScript library for creating visualizations that runs in the Adobe Flash Player.

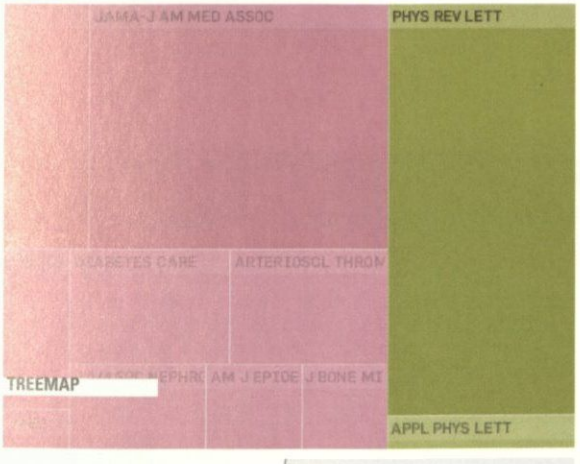
AUTHORS	Moritz Stefaner (visualization), Martin Rosvall, Jevin West, and Carl Bergstrom (Eigenfactor score)
COUNTRY	Germany and U.S.
DATE	2009
MEDIUM	Online interactive application
URL	http://well-formed.eigenfactor.org
DOMAIN	Scientific citation network
TASK	To provide an overview of information flow in science
STRUCTURE	Set of four visualizations, each with a different structure
DATA TYPE AND VISUAL ENCODING	
Categorical:	Four scientific fields: medical sciences, natural sciences, formal sciences, social sciences
Encoding:	Color: Purple, green, blue, orange
Quantitative:	Eigenfactor score
Encoding:	Radial diagram: Length of arc segment Flow diagram: Thickness of line Treemap: Area size of squared shape Map: Area size of circle
Quantitative:	In and out citation flows for each journal
Encoding:	Radial diagram: Line width and opacity Flow diagram: Not encoded Treemap: "Magnetic pins" size Map: Area size of circle
Temporal:	Five years in the dataset
Encoding:	Flow diagram: Horizontal axis

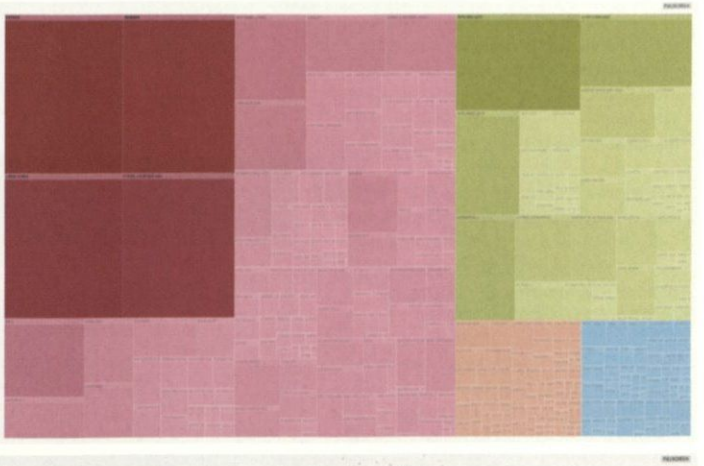

LYTIC CHEMISTRY
LUID MECHANICS
OTHER
CROP SCIENCE
& MICROBIOLOGY
EDICAL IMAGING
RADIAL DIAGRAM

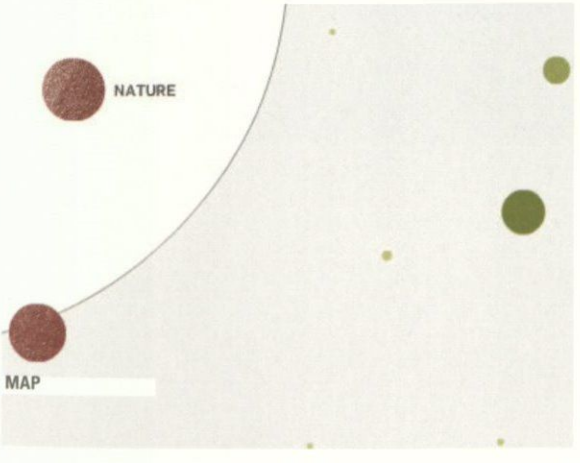


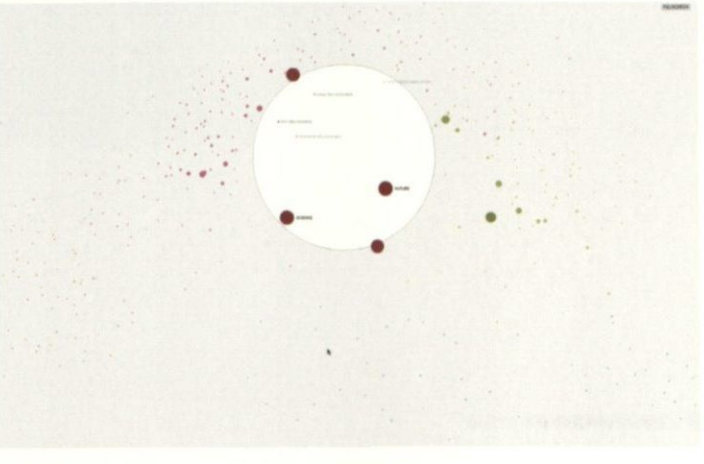

FLOW DIAGRAM




JAMA-J AM MED ASSOC
PHYS REV LETT
ARTERIOSCL THROM
TREEMAP
APPL PHYS LETT



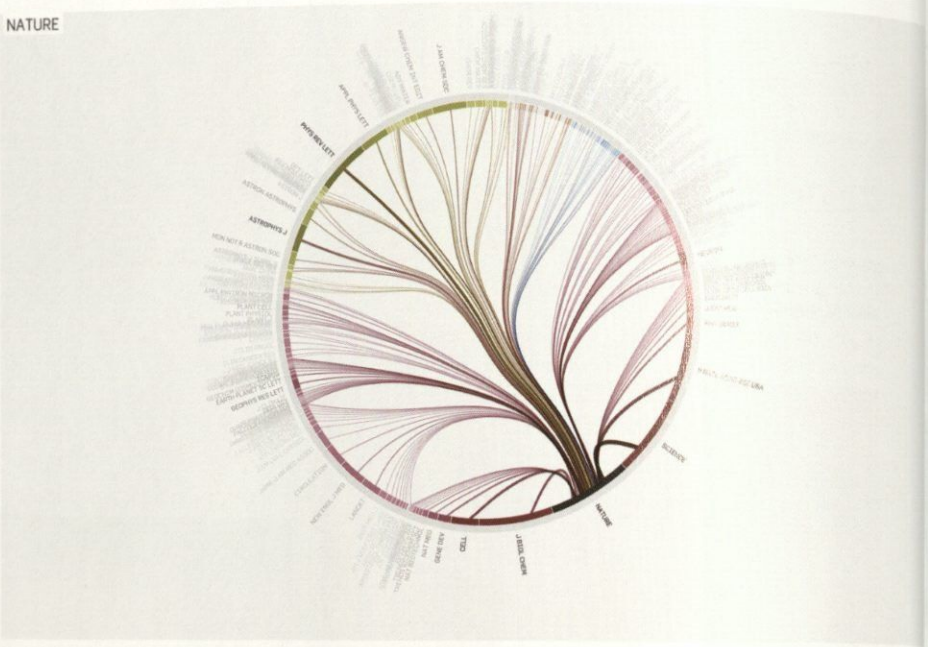

NATURE
MAP



CITATION PATTERNS

The radial diagram gives an overview of the citation network. The color scheme depicts the four main groups of journals, which is carried out through the whole set of visualizations. The outer ring portrays major fields within each of the four groups, which is further subdivided into individual journals as represented in the innermost ring. Each journal's segment is scaled by the Eigenfactor score. The citation links follow the cluster structure, using the hierarchical edge bundling technique, originally devised by Danny Holten.[22] Line width and opacity represent connection strength.

Selecting a single journal (inner ring) or a whole field (outer ring) displays all citation flow coming in or out of the selected segment. The color is based on the cluster color of the origin node.

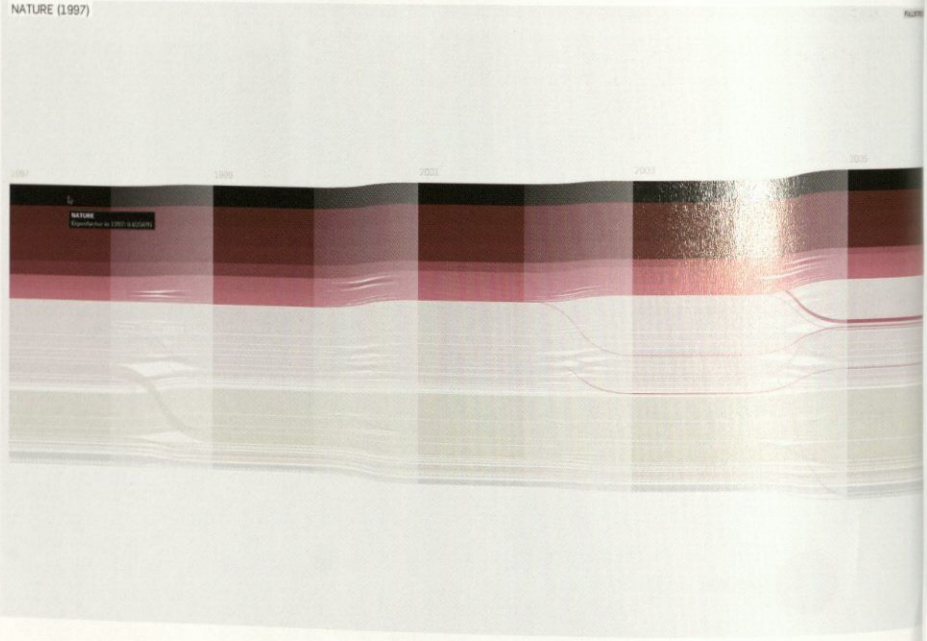


CHANGE OVER TIME

The authors call it an *alluvial* diagram, and it displays changes in the Eigenfactor score and clustering over time. It was inspired by stacked bar charts and Sankey diagrams. The latter technique is discussed in the Fineo case study later in the chapter (see page 70).

The journals are grouped vertically by their cluster structure and horizontally by year. Bars belonging to the same journal are connected. The visualization portrays five years in the dataset, each corresponding to a column: 1997, 1999, 2001, 2003, and 2005.

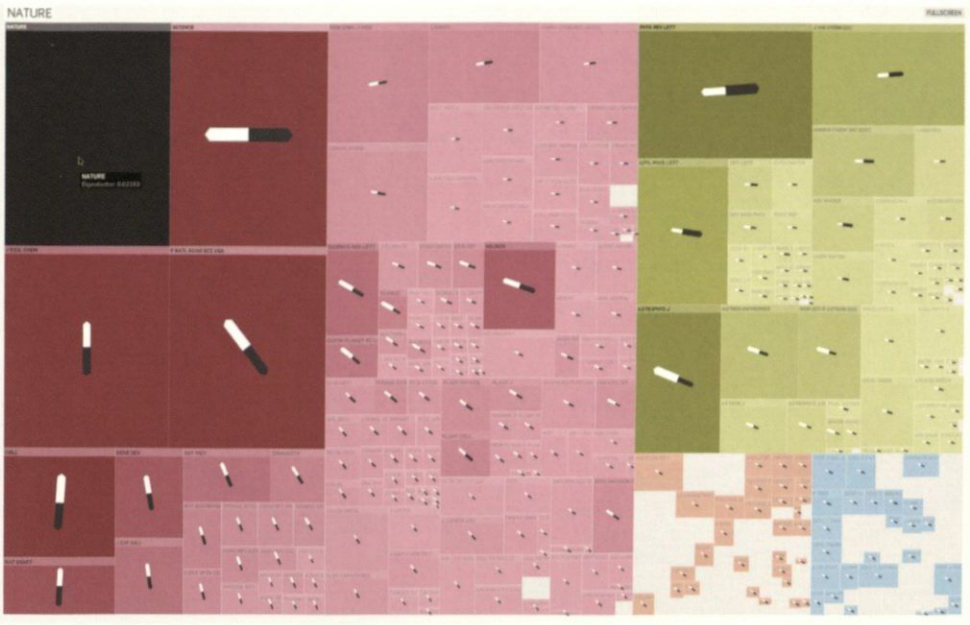
Clicking on a line highlights a journal over the years, allowing one to examine clusters the journal has been part of, track changes of influence, and determine its cluster structure.



CLUSTERING

The structure of this visualization is the squarified treemap layout algorithm, discussed in detail in the case study of chapter 1, *SmartMoney Map of the Market* (see page 30). The size of each square corresponds to that journal's Eigenfactor score.

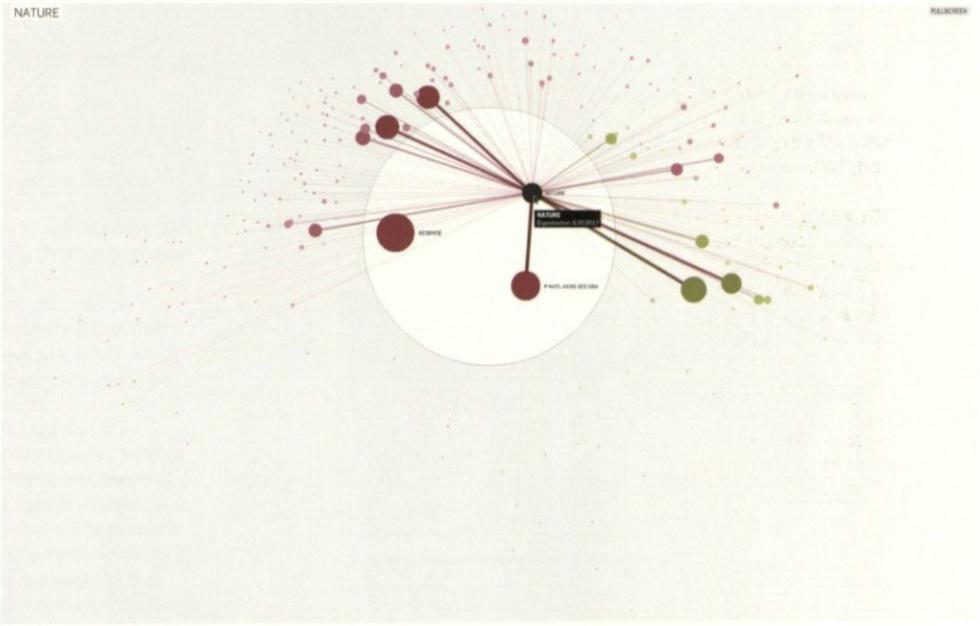
Clicking on a journal (square) displays the amount of citation flow from other journals. The flow is indicated by "magnetic pins" depicting both incoming (white arrow) and outgoing (black arrow) citation flow for any selected journal. The arrow size indicates the amount of citation flow.

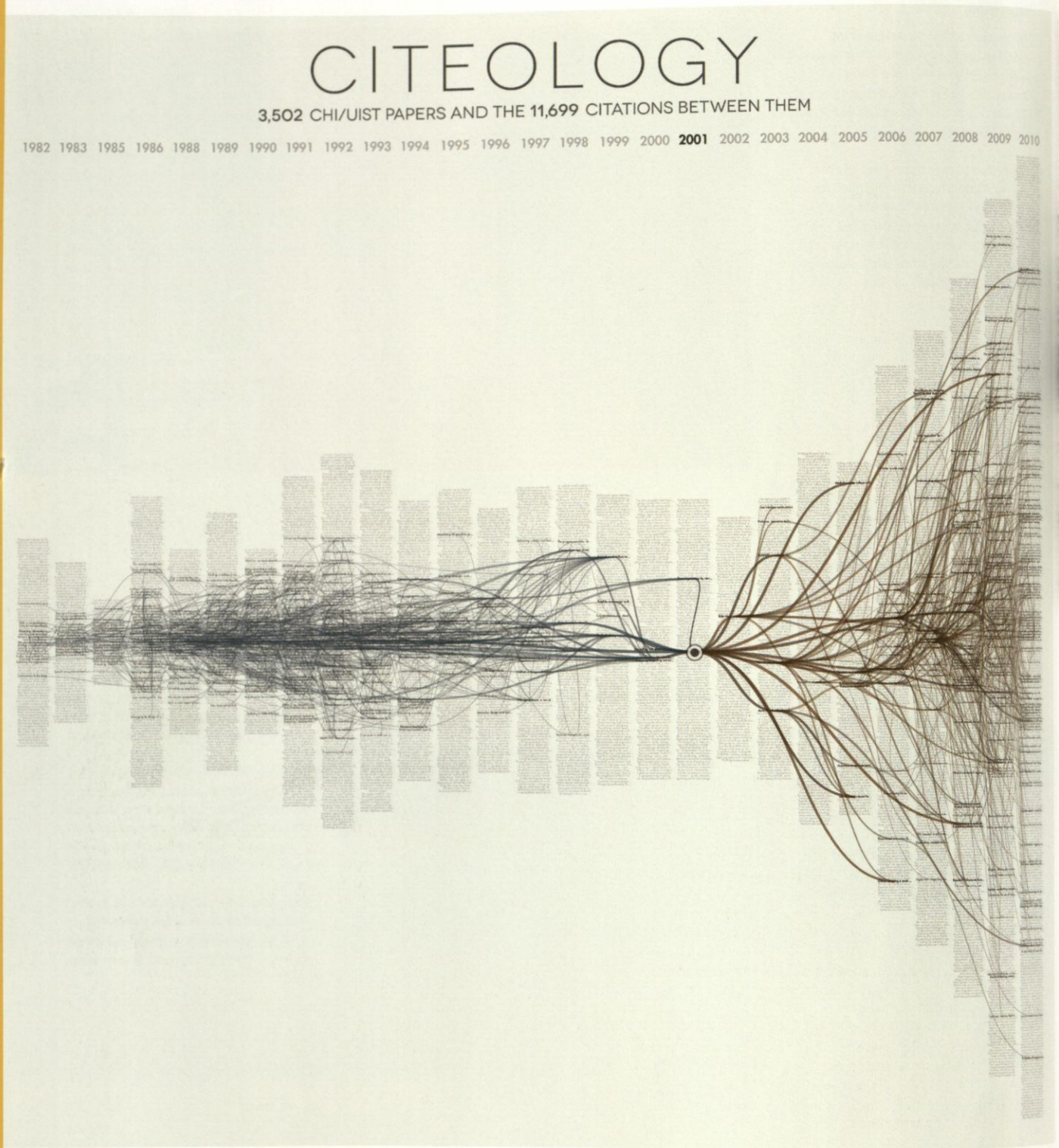


MAP

Called a map by the authors, this visualization locates journals that frequently cite each other closer together. To enlarge a part of the map for closer inspection, one can drag the white magnification lens around.

Clicking on a journal redraws the map into a force directed graph centered on that node, that is, the journal's citation network (nodes and links). The journal's area size resizes to represent the relative amount of citation flow (incoming and outgoing) with respect to the selection. When nodes are not selected, the areas are scaled by the Eigenfactor score.



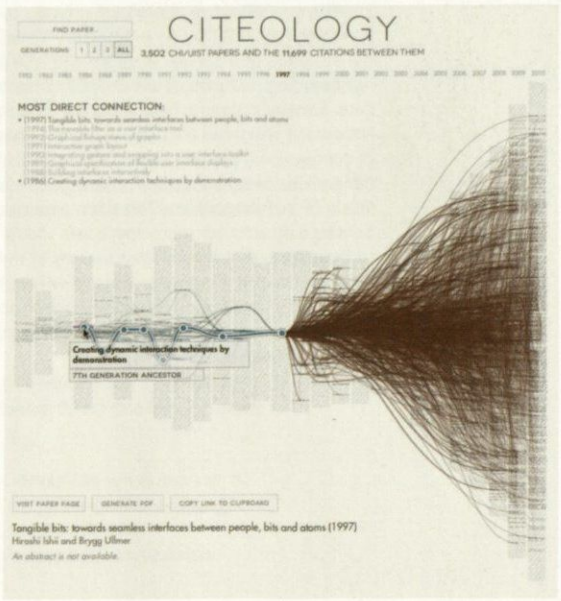

CITEOLOGY
3,502 CHI/UIST PAPERS AND THE 11,699 CITATIONS BETWEEN THEM
1982 1983 1985 1986 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Justin Matejka, Tovi Grossman, and George Fitzmaurice (Autodesk Research), Canada: "*Citeology*," 2011.

Citeology is an interactive visualization application that looks at the relationships between research publications through their use of citations. In total, 11,699 citations were made from one article to another within the collection of 3,502 papers published at two series of conferences by the Association for Computing Machinery (ACM) between the years 1982 and 2010: the Conference on Human Factors in Computing Systems (CHI) and the Symposium on User Interface Software and Technology (UIST) on Computer-Human Interaction.

Time runs horizontally and is measured in years, with the omission of 1984 and 1987, when conferences didn't occur. Papers are organized vertically by year and positioned starting at the center of each column and sorted by the frequency of citations. In other words, the papers with the largest number of citations are found at the horizontal center of the visualization. The initial twenty-five characters of papers form the lines that represent each accordingly. Because the type is too small to read on the screen, hovering over one of the lines provides the paper title. When a paper is selected, the program draws its citation network, rendering in blue connections to papers cited in the paper (descendants) and in red papers that cited it (ancestors). Thickness and opacity of the connecting lines encode age, such that lines connecting close generations are thicker and opaque in contrast to thinner and more transparent lines for further generations. The shortest path between two papers can be found by means of interactions once a paper has been selected and its *citeology* rendered.

www.autodeskresearch.com/projects/citeology



CASE STUDY

LINEAR STRUCTURE

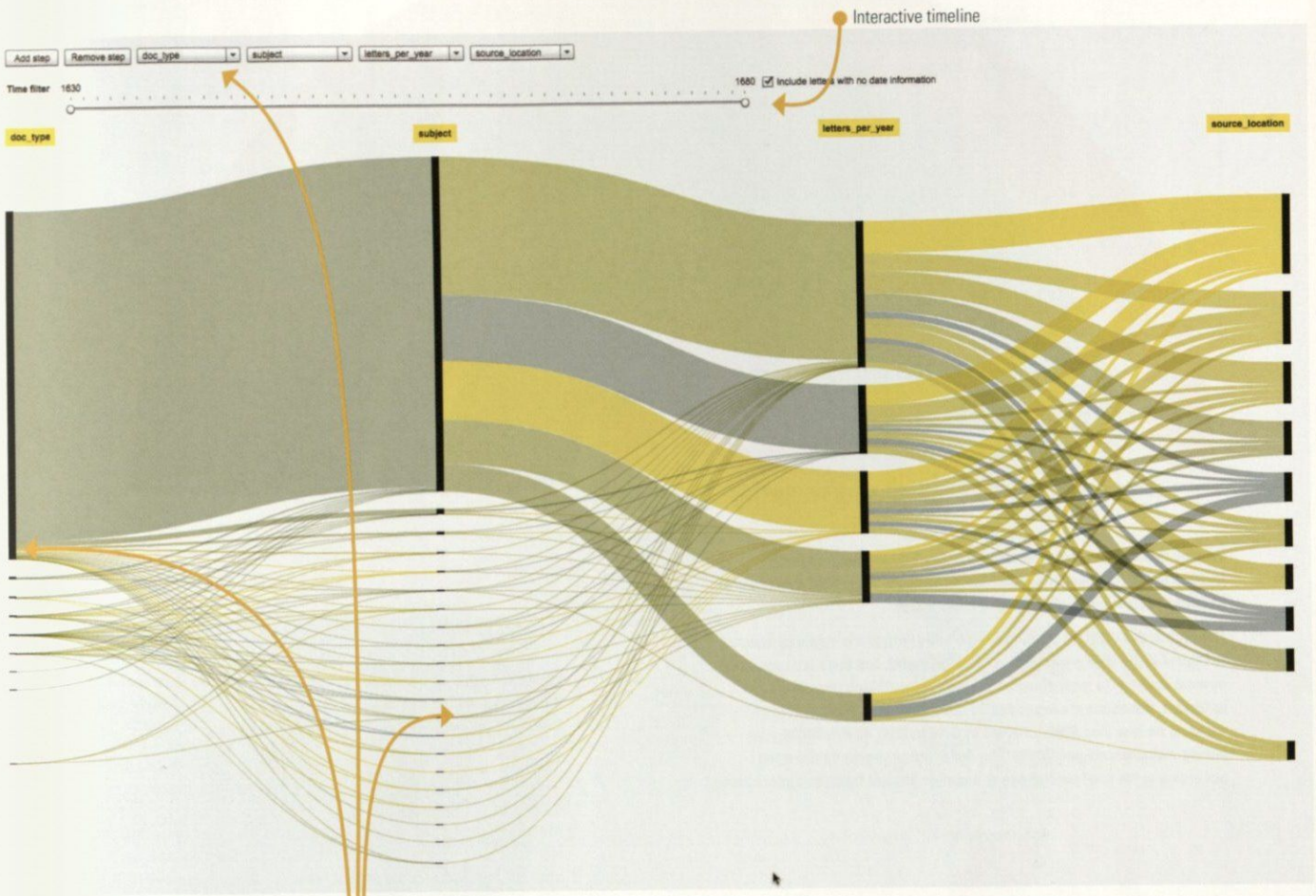
Fineo

http://fineo.densitydesign.org

First published in 1898 to depict energy flows and energy losses in a steam engine, the Sankey diagram was named after its creator, the Irish engineer Matthew H. P. R. Sankey.[23] Sankey diagrams are flow diagrams in which the widths of bands are scaled to the corresponding quantities of flow. Common examples are found in energy and financial systems, because they require understanding of the flow distribution of a phenomena.

Fineo is an interactive application created by DensityDesign Research Lab in 2010. The exploratory visualization uses the structure of a Sankey diagram with the purpose of representing relations between multidimensional categorical data. Sankey diagrams have a networklike structure, with nodes and weighted links. By using the continuous flows of connections, the tool allows easy comparison between dimensions at both local (pairs) and global (all dimensions) levels of the phenomena. The team explains, "Flows in Sankey diagrams act much more like 'rivers' (as opposed to threads) in which you lose memory of the previous steps. This can be useful in those cases in which the user is more interested in relating different data dimensions next to each other more than centering the visualization partition around a leading dimension."[24]

Bendix and Kosara devised in 2006 the Parallel Sets interactive visualization for exploring categorical data.[25] The method extends the parallel coordinates methodology by representing the set of categories along each axis, while scaling the categories according to their corresponding frequencies. Although it is similar at first glance with Parallel Sets, *Fineo* depicts data in a nonhierarchical way. In *Fineo*, axes are independent of each other, and they can be reordered to facilitate comparison between pairs of dimensions, such that one can read the visualization from all directions (left or right).

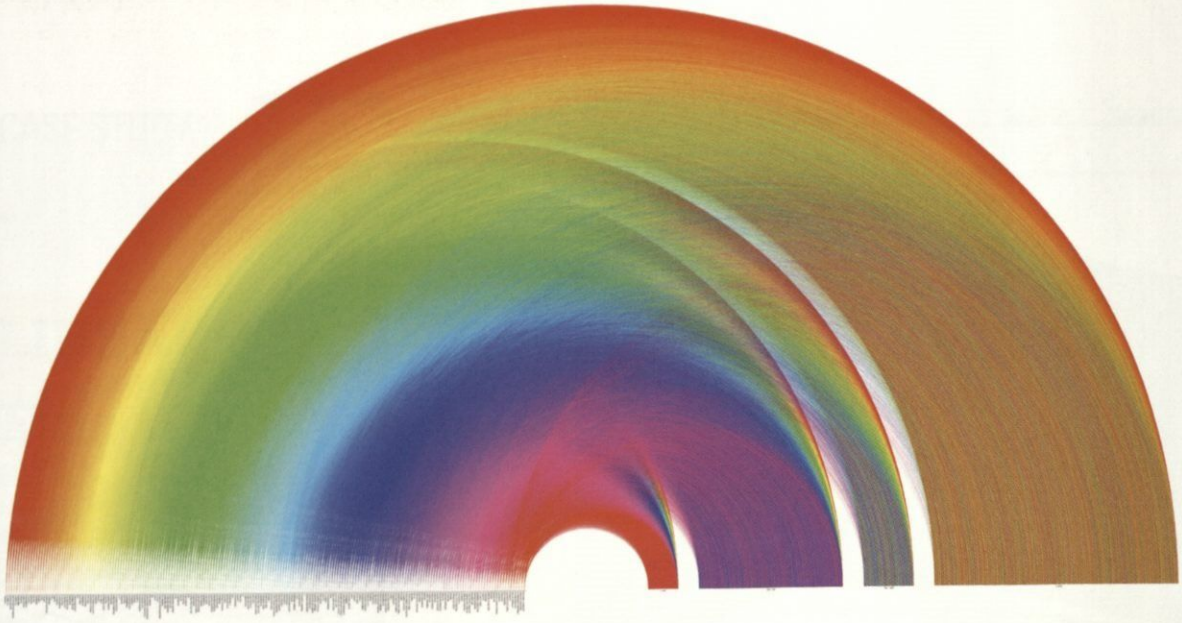


Fineo has a network structure, where nodes are individual categories grouped under a dimension, with the flow lines representing connections. Connections are grouped at every level, thus providing the width between pairs of axes.

AUTHORS	Paolo Ciuccarelli (scientific coordinator); Giorgio Caviglia, Michele Mauri, Luca Masud, Donato Ricci (researchers), at DensityDesign Research Lab, Politecnico di Milano
COUNTRY	Italy
DATE	2010
MEDIUM	Online interactive application
URL	http://fineo.densitydesign.org
DOMAIN	Categorical data
TASK	To represent relations between multidimensional categorical data
STRUCTURE	Visualization technique of continuous flow of data based on Sankey diagram structure

DATA TYPE AND VISUAL ENCODING

Categorical:	Main categorical groups. This example uses sample data from the *Republic of the Letters* project and contains seven main groups (http://fineo.densitydesign.org/mrofl/new/letters).
Encoding:	Vertical axes represent the main categories, each subdivided into subcategories
Categorical:	Subcategories
Encoding:	Vertically aligned bars (nodes), with the height defined by its frequency of occurrence
Categorical:	Connections between nodes
Encoding:	Line connecting nodes between pairs of vertical axes. Line width corresponds to the frequency of connected nodes. Color codes are categorical.
Temporal:	Years in the dataset. In this case, fifty years.
Encoding:	Interactive timeline acts as a filter in the dataset

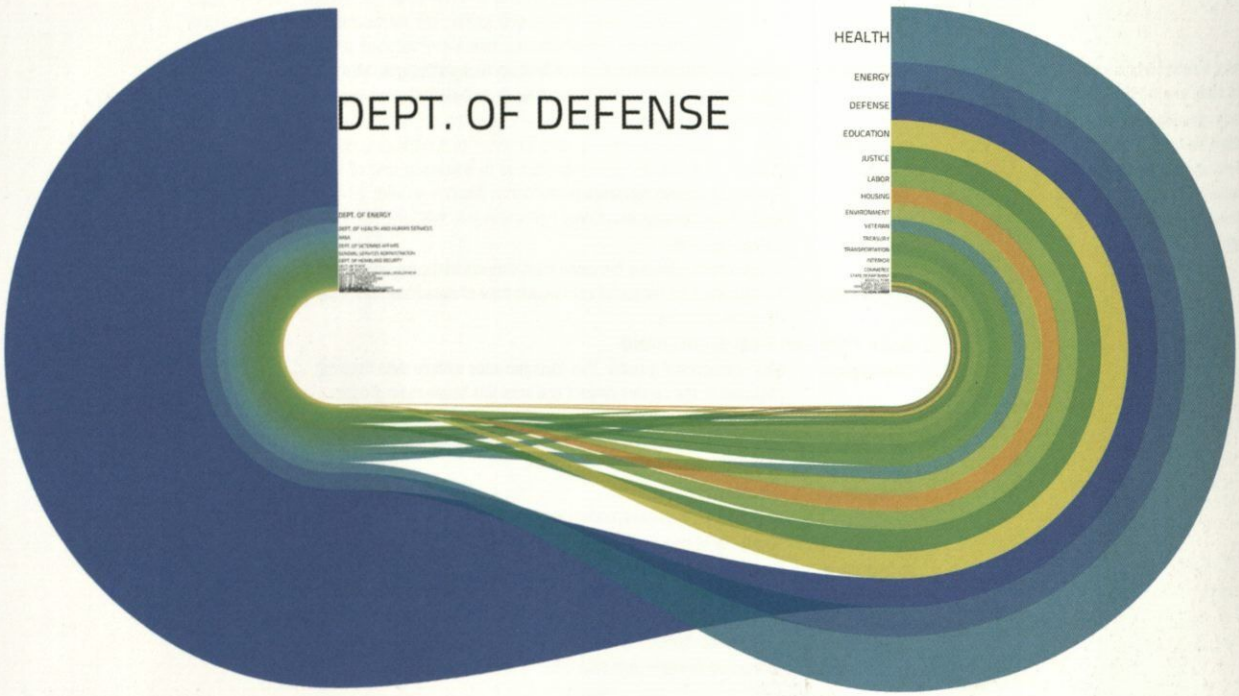


Wesley Grubbs (creative director), Nicholas Yahnke (programmer), Mladen Balog (concept artist) at Pitch Interactive, U.S.: "2008 Presidential Candidate Donations: Job Titles of Donors," 2008.

The arcs in this diagram connect the job titles (left) to the amounts donated to Obama in the 2008 presidential campaign (right). Job titles are organized by most common to least common among the top 250 job titles of donors to Obama. Donations are organized by dollar amounts, with the first group standing for less than $100, followed by $100 to $500, $500 to $1000, and ending with amounts over $1000. The dollar group segments are sized according to the total percentage of donation amount from the donors listed.

Wesley Grubbs (creative director), Nicholas Yahnke (programmer), Mladen Balog (concept artist) at Pitch Interactive, U.S.: "US Federal Contract Spending in 2009 vs. Agency Related Media Coverage," 2010.

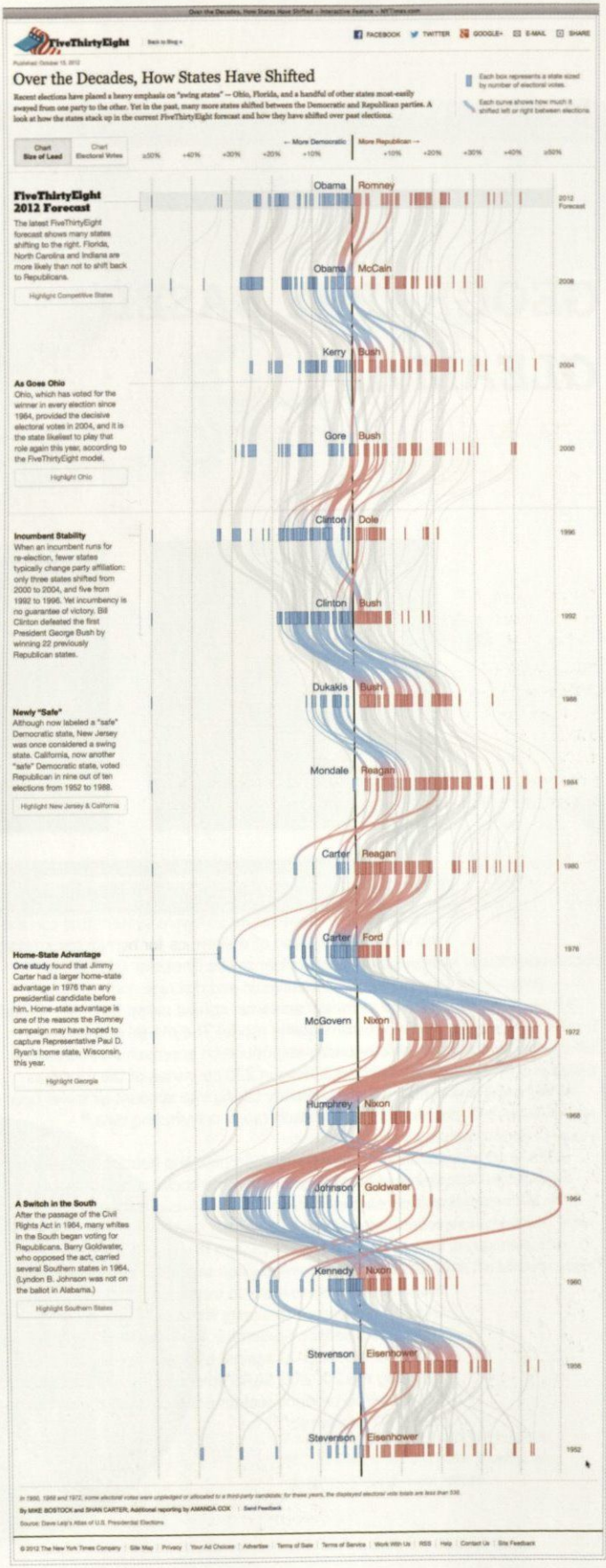
The graphic plots U.S. federal agency spending in 2009 against media coverage of those agencies in the same year. Each agency is represented by a stripe proportional to its budget presence. The graphic reveals that there is a dramatic mismatch between what American taxes fund and which issues occupy national discourse. It is clear for example, that defense spending accounts for the majority of the federal budget, almost 70%.



Mike Bostock, Shan Carter, and additional reporting by Amanda Cox (*New York Times*), U.S.: "Over the Decades, How States Have Shifted," 2012.

The flow diagram visualizes how American states have shifted parties over the years, from 1952 to 2012. Time is organized vertically with each row representing an election year. The horizontal axis depicts the size of the parties. Each line represents a state. For example, one can select a line and see the path taken by a state across elections, and whether it has changed, and how much. Note the information on the left-hand side, that provides contextual information and helps the viewer further interpret the diagram.

www.nytimes.com/interactive/2012/10/15/us/politics/swing-history.html

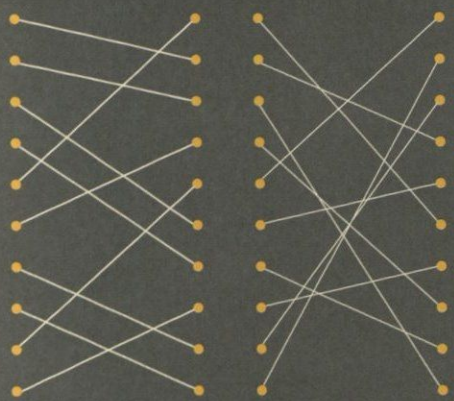


Common Fate

The common fate principle is the tendency to group elements that are moving in the same direction. For example, we tend to perceive the six lines below as forming different groups:

\ \ \ \ \ \ = 1 group
\ \ \ / / / = 2 groups

Parallel lines are easier to perceive and will be grouped into a unit, whereas nonparallel lines are perceived individually.

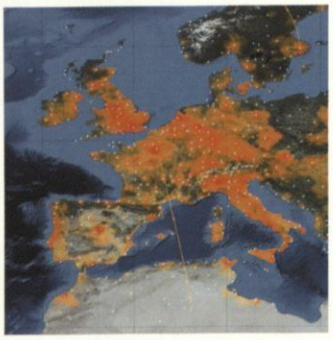


CASE STUDY

GEOGRAPHY BASED

GLEAMviz

www.GLEAMviz.org



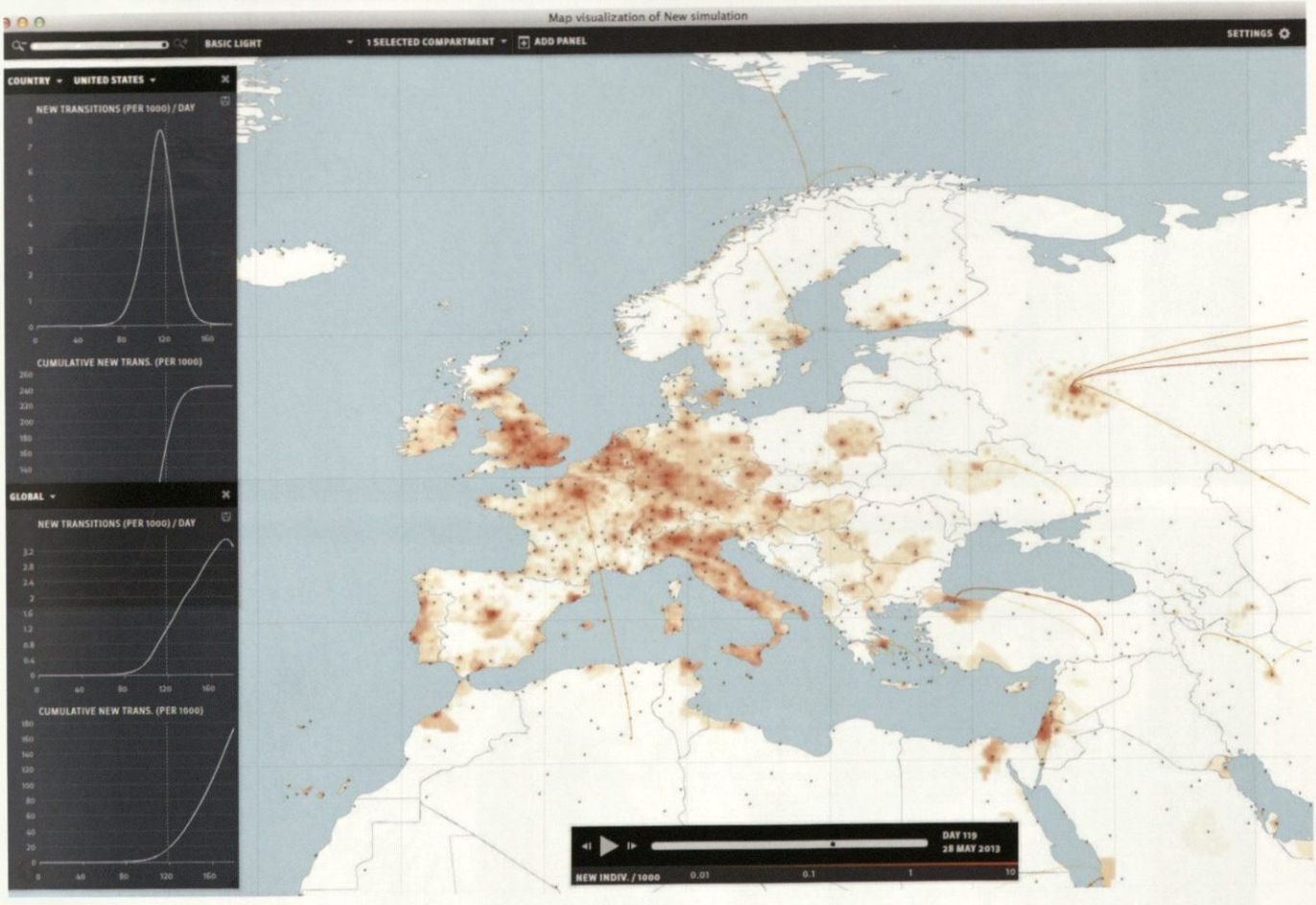
GLEAMviz is a client-server software system that can model the worldwide spread of epidemics for human transmissible diseases such as influenza-like illnesses. GLEAMviz makes use of a stochastic and discrete computational scheme to model epidemic spread called GLEAM—Global Epidemic and Mobility model. The model is based on a geo-referenced metapopulation approach that considers 3,362 subpopulations in 220 countries of the world, as well as human mobility taking into account air travel flow connections and short-range commuting data.[26]

Epidemic forecasts are complex and need to consider a series of parameters within the social context. Vespignani and colleagues explain, "GLEAM uses real-world data covering the distribution of the worldwide population, their interactions and journeys, and the spatial structure and volumes of national and international air traffic. By combining these datasets with realistic models of infection dynamics, GLEAM can deliver forecasts for the spreading pattern of infectious diseases epidemics. We have thoroughly tested and validated GLEAM against historical epidemic outbreaks including the 2002/03 SARS epidemic. In 2009, GLEAM has been used to produce real-time forecasts of the unfolding of the H1N1 pandemic."[27]

GLEAMviz offers three types of visualization:

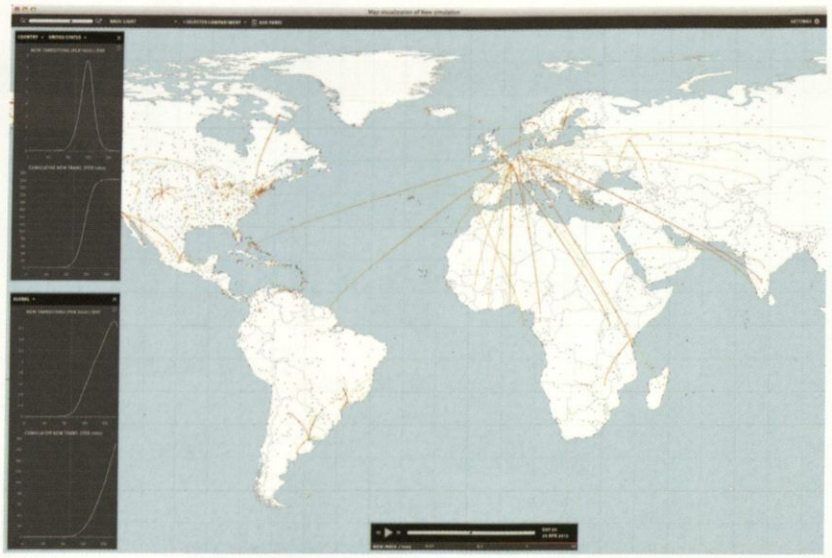
- A 2-D map depicting the spread of the infection with charts showing the number of new cases at various levels of detail.
- A 3-D globe with a concise overview of the spread.
- Geographic and concentric views by SPaTO visual explorer depicting how the structure of the airport network influences the notion of distance. The outputs remap all the transportation hubs according to the time it takes for the infection to reach them from the moment of outbreak.

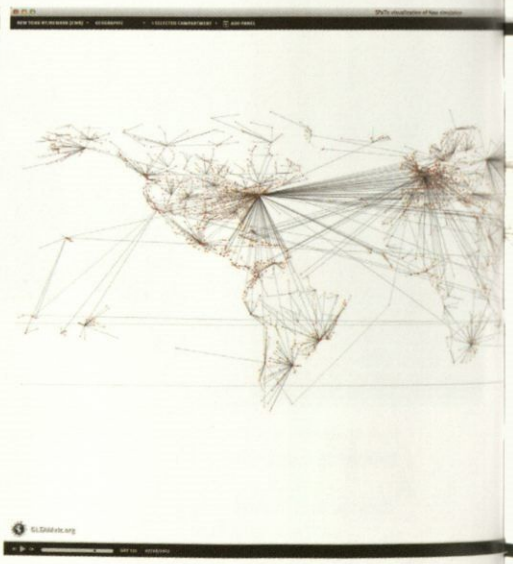
GLEAMviz, which is publicly available for download, allows setting up and executing simulations, and retrieving and visualizing the results. It was developed by an international team lead by Alessandro Vespignani (team coordinator) and Vittoria Colizza. It is hosted at three institutions: College of Computer and Information Sciences and Department of Health Sciences, Northeastern University, Boston, MA, USA; Complex Networks & Systems Group, ISI Foundation, Turin, Italy; and INSERM, Unite Mixte de Recherche 707, Paris, France.

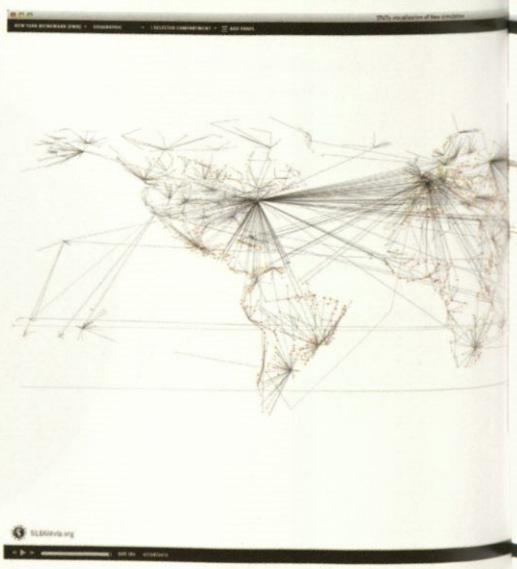


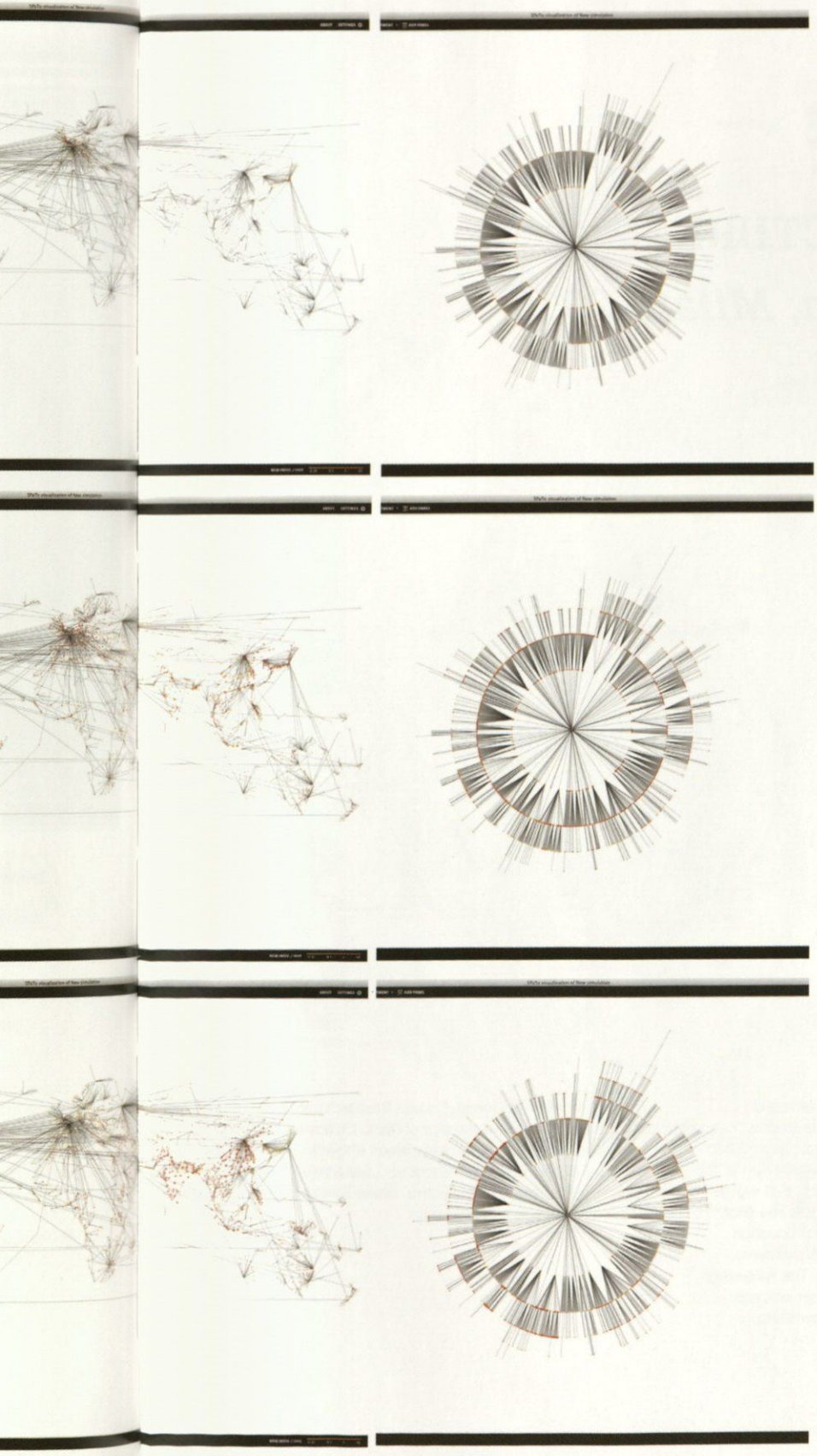
The image shows one of the possible visualization outputs in GLEAMviz: a 2-D geo-temporal evolution of the epidemic. The map shows the state of the epidemic on a particular day. Infected population cells are color coded according to the number of new cases of the quantity that is being displayed. Detailed information is provided on demand by clicking on a city. The evolution of the epidemic can be viewed as an animation by using the play button at the bottom of the interface. The two sets of charts (on the left) depict the incidence curve and the cumulative size of the epidemics for selectable areas of interest. There are three options for map backgrounds: Blue Marble map by NASA's Earth Observatory, a dark or white solid color.

Among the types of visualization available at GLEAMviz is this 3-D globe showing an overview of the disease spread.





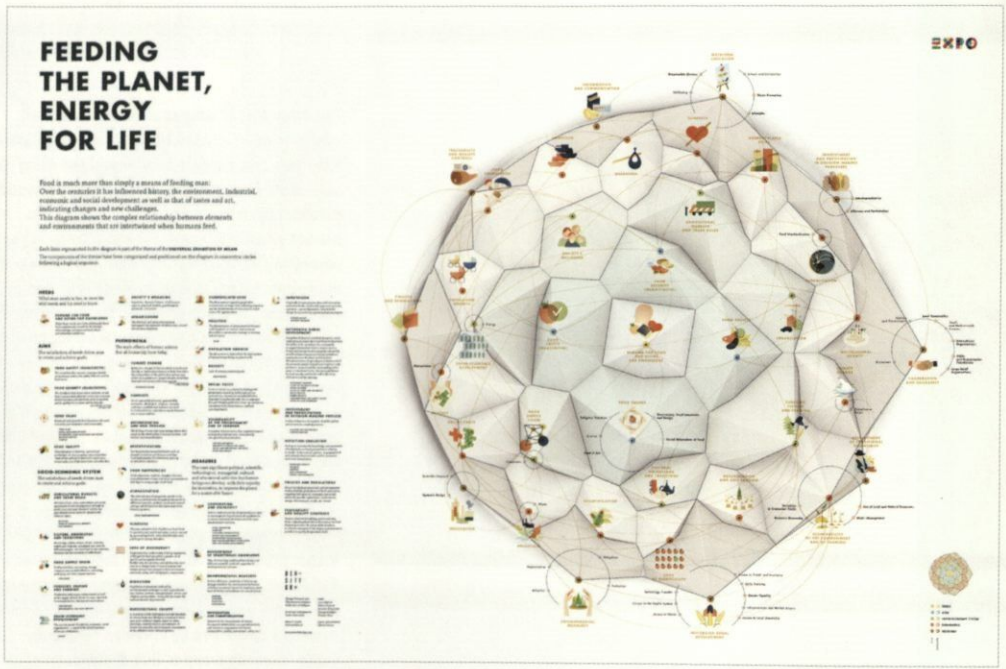




The three sets of images show the temporal evolution of an Influenza Like Illness (ILI) started in New York City simulated and visualized using the GLEAMviz simulator. The model uses real-world data on population and mobility networks to predict when and where people will interact and potentially transmit the infection. The sets show the result of disease spread for days 90 (top), 130 (center), and 180 (bottom). The left image in each set shows the geo-temporal evolution of the epidemic for each particular day. The arrows show the spread of infection. The color of each census area shows the local number of transitions into the infected compartment (incidence). The two black chart panels on the left show the incidence and total number of cases in the United States (top) and in the globe (bottom) as a function of time.
The remaining images are renditions from *SPaTo Visual Explorer*, a visualization tool integrated into GLEAMviz. *SPaTo Visual Explorer* is an interactive tool for the visualization and exploration of complex networks developed by Christian Thiemann in the research group of Dirk Brockmann at Northwestern University. The system provides two views: geographic and concentric. It uses a radial distance corresponding to "effective" network distance, that is, the shortest-path distance to the central node. As Thiemann explains, "By reducing a network to the shortest-path tree of a selected root node, we obtain a local but simpler view of the network that can be easily visualized. With the ability to quickly change the root node, the program allows us to explore the network from different perspectives."[28]

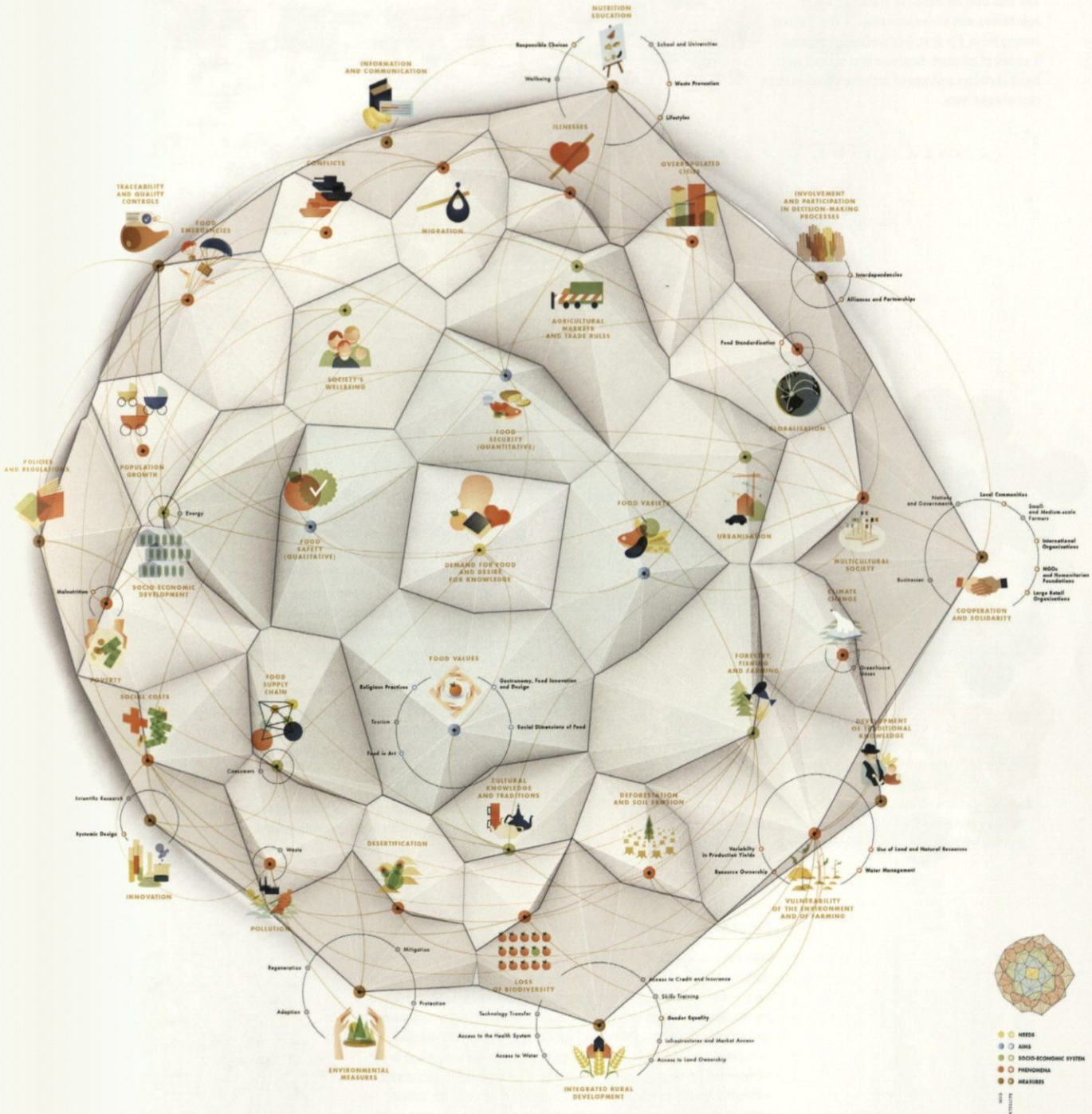
COMMUNITY STRUCTURE

Universal Exposition, Milan, 2015



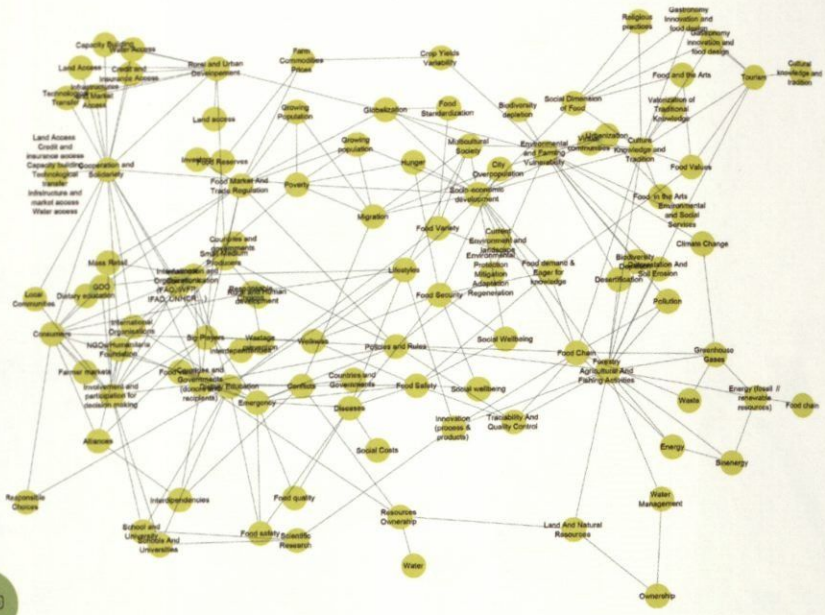
The city of Milan, Italy, will host the 2015 Universal Exposition around the topic of food with the theme "Feeding the Planet, Energy for Life." The organizers of Expo 2015 invited the Italian research laboratory DensityDesign at the Politecnico di Milano to devise a visualization that would communicate the topic to a general audience. The final poster depicts relationships between food production and consumption, social and environmental concerns, and technological and sustainability issues. The following spread provides an explanation of the design process, from the technical graph to the design of symbols.

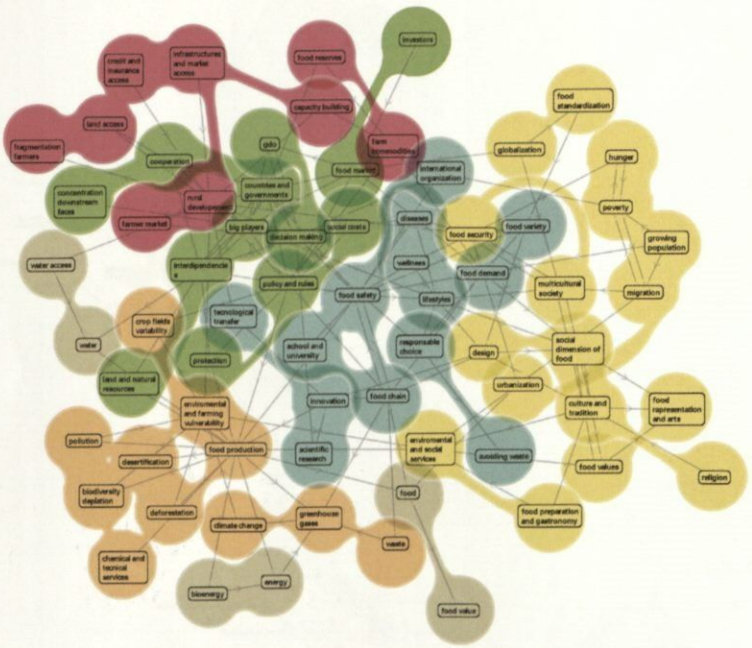
The design team at DensityDesign Research Lab, Politecnico di Milano consisted of Paolo Ciuccarelli (Scientific Coordinator), Michele Mauri (Project Leader), Giorgio Caviglia, Lorenzo Fernandez, Luca Masud, Mario Porpora, and Donato Ricci (Team). Gloria Zavatta took part in the theme development.


NUTRITION EDUCATION
Responsible Choices
School and Universities
Wellbeing
Waste Prevention
Lifestyles
INFORMATION AND COMMUNICATION
ILLNESSES
CONFLICTS
OVERPOPULATED CITIES
TRACEABILITY AND QUALITY CONTROLS
FOOD EMERGENCIES
MIGRATION
INVOLVEMENT AND PARTICIPATION IN DECISION-MAKING PROCESSES
Interdependencies
Alliances and Partnerships
AGRICULTURAL MARKETS AND TRADE RULES
Food Standardization
SOCIETY'S WELLBEING
GLOBALISATION
FOOD SECURITY (QUANTITATIVE)
POLICIES AND REGULATIONS
POPULATION GROWTH
FOOD VARIETY
URBANISATION
Local Communities
Nations and Governments
Small and Medium-scale Farmers
International Organizations
FOOD SAFETY (QUALITATIVE)
Energy
DEMAND FOR FOOD AND DESIRE FOR KNOWLEDGE
MULTICULTURAL SOCIETY
NGOs and Humanitarian Foundations
Businesses
Large Retail Organizations
SOCIO-ECONOMIC DEVELOPMENT
Malnutrition
CLIMATE CHANGE
COOPERATION AND SOLIDARITY
Greenhouse Gases
FOOD VALUES
FORESTRY, FISHING AND FARMING
POVERTY
FOOD SUPPLY CHAIN
Religious Practices
Gastronomy, Food Innovation and Design
SOCIAL COSTS
Tourism
Social Dimensions of Food
Food in Art
DEVELOPMENT OF TRADITIONAL KNOWLEDGE
Customers
CULTURAL KNOWLEDGE AND TRADITIONS
DEFORESTATION AND SOIL EROSION
Scientific Research
Systemic Design
DESERTIFICATION
Variability in Production Yields
Use of Land and Natural Resources
Resource Ownership
Water Management
Waste
INNOVATION
VULNERABILITY OF THE ENVIRONMENT AND OF FARMING
POLLUTION
Mitigation
Regeneration
LOSS OF BIODIVERSITY
Access to Credit and Insurance
Skills Training
Adaptation
Protection
Gender Equality
Technology Transfer
Infrastructures and Market Access
Access to the Health System
Access to Water
Access to Land Ownership
ENVIRONMENTAL MEASURES
INTEGRATED RURAL DEVELOPMENT
NEEDS
AIMS
SOCIO-ECONOMIC SYSTEM
PHENOMENA
MEASURES
SIDE
DATASETS

THE NETWORK

The first step involved the identification of subthemes and the examination of interactions among them. For that, DensityDesign created a series of network diagrams that started with hand sketches and ended with the digital version reproduced here.





CLUSTERING

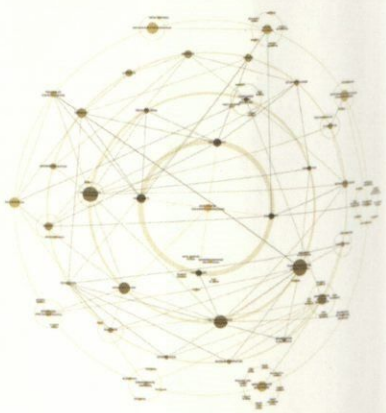
After creating the relational structure, they grouped the subthemes into five categories:

- Needs: The needs of humans to live and to meet vital needs, and our needs to know;
- Aims: The satisfaction of our needs drives humans to create products and to achieve goals;
- Socioeconomic system: The productive, social, economic environment;
- Phenomena: The main effects of human actions that all humanity faces today;
- Measures: The most significant political, scientific, technological, managerial, cultural, and educational activities that human beings can develop, with their capacity for innovation, to improve the planet for a sustainable future.

Each color stands for one of five categories structuring the subnetworks within the main theme.

THE GRAPHICAL LAYOUT

With the subnetworks organized into five categorical groups, they started studying the best visual representation with which to communicate the story to the general audience. In the series of representations, we see the iterations of the design process toward a layout that would maintain the complexities of the theme without the technicality of the initial network diagram. In this process, each subtheme became a hub in the network with their connected nodes. More important, the subnetwork within the Needs category was centralized in relation to the whole network and surrounded by the other four groups. The last step in the graphical structure was the introduction of the landscape metaphor. Using Delaunay triangulation, they assigned areas to the groups so as to depict the network as a Voronoi diagram. Each node is represented as a mountain peak, with its height provided by the number of connections (the node degree).



PICTOGRAM SYSTEM

The last stage involved the creation of a series of pictograms to convey each theme. The color code stands for the five main themes. Each hub is represented by a diamond shape inside a circle of its theme color, and other nodes are represented by outlined circles, again color coded for themes.