Robert Head
EMDA 331 / ART 507
Fall 2016

Visualizing Complexity: Mapping Patterns of Information
By Manuel Lima
Chapter 2: From Trees to Networks
Chapter 5: The Syntax of a New Language

As a software developer, my primary task is to come up with useful abstractions – mental and technical models that reflect, as accurately and usefully as possible, the underlying realities of the problem domain. How we choose to model data, how we think about the breakdown of a problem, deeply impacts the ease with which we can draw valid conclusions and provide a useful tool to interact effectively.

During some of my early work at Amazon.com, it became clear that deep hierarchies fail to adequately model the real world or to provide a useful tool at high levels of scale. Consider this. How do you organize all of the books in the English language? Or all the products in a store? We can try to create a tree structure in which we impose layers of categorization. But where do we start? By topic? By target market? By medium? By price? By all of them in every possible combination?

As Amazon quickly learned, a singly rooted hierarchy of millions of products is both difficult to create and difficult to navigate. Worse, it does not ensure that a customer will know where to find something. Should a camping cookbook be found in the camping section or the kitchen section? For a digital company, the cost of duplicate placement is irrelevant, so the answer can be both.

Amazon's solutions, which survive to the present day, took us beyond a hierarchical ‘browse’ experience of the tree data structure. Instead, the site provides a rich web of searchable dimensions. Multiple navigable, shallow (but interrelated) hierarchies, some curated by humans and others grouped by data analysis, are available and combined with search term results. The rich dataset underneath allows for a seamless browse-search-filter combination of utility.

In Manuel Lima's book, *Visualizing Complexity: Mapping Patterns of Information*, new patterns of modeling data are explored that empower us to think in new ways. In the second chapter, he encourages the reader to move beyond the traditional tree structure and develop sets of nodes with richer possibilities. Much as Amazon embraced multiple threads of connection between works, Lima envisions a new grammar for a visual language of interlinked networks.

In the world of data models, a tree is a structure in which each data member has one parent and any number of children. Trees are, essentially, "singularly rooted" categorization. The Dewey Decimal System is a well-known example, with categories moving from general to more specific as the branches radiate outward and decimal places move to the right.

As models, however, trees impose an artificially rigid, hierarchical order on systems that might in fact be more complex or not intrinsically ordered. Trees embrace a philosophy of finalism, in which the hierarchy implies a flow in one direction to an inevitable end result. Trees also project an expression of essentialism, in which things are thought to have a fixed set of underlying properties defined by a node's position within the data structure. Tree structures are, then, defective in describing the complexities of a wholism by implying an artificial centralism, a uni-directional finalism, and a stagnant essentialism.

In place of tree structures, perhaps we should prefer the semilattice or network, which allows for multiple connections between nodes. In keeping with the botanical metaphors, a rhizome is an exploratory structure and reaches out in many ways at once without a sense of center. The rhizome as a data structure is described as flexibly ordered and defined by a system of interconnected states. This type of structure more ably reflects the stage of data science that we have achieved in recent decades.

Historically, science has evolved through several stages. At first, we examined simple problems with a single or a small number of variables. Beginning in the late Nineteenth century, we began to explore gestalts and integrated systems – problems of disorganized complexity. Recently, as data collection and processing have grown exponentially, we have arrived at an understanding of the interconnected and interdependent variables of organized complexity.

The application of new thinking about interrelationships is reflected in many endeavors, from urban planning to evolutionary biology, brain science, and technology design.

Categorization was intrinsic to mid-century urban planning. Separate residential, commercial, and industrial districts created a sense of disconnection, compartmentalization, and the pain of time lost to commuting. Rebellious urban planners eventually began to insist on an integrative approach, permeable spaces, and mixed-use designs.

In the life sciences, an ongoing effort has been made to categorize the world’s creatures. The classical philosophers and early scientists grouped based on shared characteristics. This model was replaced, starting with Darwin, with a tree structure that traces the evolutionary lineage of each creature back through common ancestors–a huge family tree. While the power and clarity of this model is undeniable, it doesn’t tell the full story. Recently, microbiologists have discovered horizontal gene transfer, not through descent, but through incorporation of material from microbes. A ‘net of life’, perhaps, is now necessary to more accurately illustrate the known flow of genetic material.

In the study of the brain, the ancient world associated facial features with character (physiognomy) and the nineteenth century practice of phrenology associated traits with the shape of the skull and presumed 'organs' within the brain. The reality of brain function, however, is that the neurons are highly integrated and there are not well-defined loci for various functions. Modern neuroscience provides us with a model of the brain that more accurately describes billions of nodes (neurons), each of which has potentially thousands of connections. Areas of function, such as visual processing, logic, or memory, are not tightly localized, but distributed.

The wisdom of this distributed design has been applied to instruments of human creation. For example, the internet architecture was the result of a project to build a network with high survivability and resistance to attack. The result is the fully distributed, rather than a centralized or finitely decentralized, architecture of the internet. This distribution network made a natural home for interconnected data, the "ubiquitous datasphere" known as the World Wide Web.

In the first decade of the Web, there was widespread use of tree structures. ‘Browsing’ tree structures was one of the primary means of navigation. Sites provided menus of a site’s content pages. Portals such as Yahoo and Open Directory even attempted to create hierarchies of all significant sites on the Web.

The emergence of what became known as “Web 2.0” in the early 2000s began a trend toward democratization such as user-driven content and categorization. One technique was using unstructured user tagging to create a folksonomy of metadata. A particular style of data visualization known as a tag cloud was even in fashion for a few years. The leaderless democratization of data generation has had many successes (wikipedia, for example). Tagging has also, over the years, shown the inherent limitations of being open and unstructured (disparate synonyms, defaceability, etc.).

These new explorations of relationships and metadata can and should form the basis of new types of data visualizations, such as radial convergences, area groupings, and arc diagrams. In the fifth chapter of Lima’s book, he presents dozens of examples of network data visualizations. These graphs all go beyond tree-structured data, but range widely in utility, comprehensibility, and aesthetic value. Wesley Grubbs and Nick Yahnke’s *Patterns in Oscar Movies* (2007) is a particularly compelling segmented radial convergence diagram. The topic is easily understood and of popular interest. The data is partially hierarchical, with nodes falling into three categories: directors, Oscar winners, and non-winning actors. The choice to place nodes in these categories as concentric circles strengthens our ability to comprehend. The connections, however, are neither hierarchical nor singly rooted. The choice of a single color of gentle bezier curves with clear programmatic consistency conveys large amounts of information in a way that manages to be aesthetically pleasing and simultaneously understated.

Anyone responsible for managing large amounts of data, such as e-commerce information or user-generated content, can testify to the need for tools that can illuminate relationships across large datasets. Visualizations of nodes networked with multiple connections are a powerful and effective new tool for the age of agile responsivity to big data and rapid change. As our data vocabulary expands, so do our visual grammars.