

Treemaps as a Tool for Social Network Analysis^{*}

CASOS Technical Report

Terrill L. Frantz & Kathleen M. Carley

September 2005

CMU-ISRI-05-118

Carnegie Mellon University

School of Computer Science

ISRI - Institute for Software Research International

CASOS - Center for Computational Analysis of Social and Organizational Systems

Abstract

We apply treemap technology in the analysis of large, social network datasets—principally for examining network subgroups. A treemap is graphically-based information and exploration tool which is used in diverse fields such as computer science, finance, and human-gene research; from our experience, we find treemaps also useful in the social network analysis setting. A treemap represents hierarchical and categorical data in a mosaic form containing embedded, rectangular shapes, where the size of each shape is germane. Treemap displays are especially helpful when examining data in an interactive mode (as opposed to a static or printed form). We have found that treemaps are a powerful tool for exploring large social-networks, particularly during the exploratory data analysis phase. Their use quickly leads to a thorough perspective of the holistic characteristics of the network and to easier identification of significant subgroups; both of these perspectives may otherwise remain hidden using traditional visualization techniques. In this report, we introduce treemap technology, first broadly, then, specifically how it can be applied to social network analysis. We also show how we have actually applied treemaps to an interactive study of a large, real-world dataset. As a result of our experiences, we encourage other social network analysts to consider applying treemaps in their work.

^{*} This work was supported by the Department of Defense and National Science Foundation under MKIDS program (IIS0218466) and the Office of Naval Research under Dynamic Network Analysis program (N00014-02-1-0973). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense, the National Science Foundation, the Office of Naval Research, or the U.S. government.

Keywords: social network analysis, large datasets, data visualization, treemaps, data exploration

Table of contents

I.	Index of Tables	6
II.	Index of Figures	6
1.	Motivation.....	6
2.	Introduction to Treemap Technology	6
2.1	Treemaps: At a Glance.....	7
2.2	Characteristic Features.....	8
2.3	An Illustrative Example	10
2.4	Variants of the Core Treemap Technology.....	11
2.5	Application in Various Fields	12
2.6	Software and More Information.....	12
3.	Application to Social Network Analysis.....	12
3.1	Treemaps in the Social Network Setting	13
3.2	Examining Subgroups.....	14
4.	Demonstration of Exploring a Large Social Network	16
4.1	Using Traditional Diagram	16
4.2	Using Treemaps	18
4.3	The Overall Experience	20
5.	Conclusions and Next Steps.....	21
6.	References.....	21

I. Index of Tables

Table 1. Features Comparison of Visualization Techniques.....	7
---	---

II. Index of Figures

Fig. 1. A Treemap of a Road Maintenance Record.....	7
Fig. 2. A Treemap of a Computer Disk Directory: File Size and Directory Hierarchy.....	9
Fig. 3. Results from Drill-Down of Prior Image (Fig. 2): Treemap of the “Help” Subdirectory.....	10
Fig. 4. Node-Link Diagram of a Large Social Network Dataset: The Enron Email Corpus.....	15
Fig. 5. Drill-Down of a Node-Link Diagram: The Enron Email Corpus.....	16
Fig. 6. Traditional Histogram of Node Degrees: The Enron Email Corpus.....	16
Fig. 7. Treemap: Actor Node Degrees, by Title	17
Fig. 8. Treemap: Actor Node Degrees, by Title, Year, & Month.....	17
Fig. 9. Treemap: Node Degrees, by Year & Month.....	18
Fig. 10. Treemap: 2001 Node Degrees, by Month.....	18
Fig. 11. Treemap: Oct. 2001 Node Degrees, by Name.....	19
Fig. 12. Treemap: Oct 2001 Node Degrees, by Title, & Name.....	19

1. Motivation

It has become evident that exploring social network data is an increasing challenge as datasets have recently become quite large. Today, the typical size of a social network dataset being studied has increased from a mere handful of actors, to millions; thanks to the advancement of data collection techniques. Consequently, creating a graphic representation of the data often results in exceeding the practical and usability limits of traditional visualization tools. In the past, the convention has been to visualize network data in the form of a node-link diagram (a sociogram), an adjacency matrix, or other common statistical diagrams: like bar and pie charts. For certain tasks, these traditional tools have become, practically, outmoded.

We recently conducted a social network study of a large email corpus that consisted of several hundred nodes. Right away, we faced the limitation of using traditional network visualization tools—for example, the node-link picture looked like countless dots superimposed on a gaudy background. It was obvious; the size of the network rendered the traditional node-link diagram virtually *ineffective*. Little, if any, useful information could be gleaned from the pretty and colorful, but appallingly hazy pictures. We presume it likely that other researchers are also facing this same problem.

Visual exploration is certainly an essential and crucial activity in any research domain. The Exploratory Data Analysis (EDA) (Tukey, 1977) of *social network* data is undoubtedly no exception. In the past, EDA has been a requisite activity for social network analysis and will likely remain so in the future (Freeman, 2000a). Without doubt, a visual representation of data is a decisive adjunct to an analyst's thorough comprehension of a social network; in addition, *quality* imagery is an asset to analysts when communicating their findings to others (Freeman, 2000b). In the present day, however, analysts *must* find tools that better meet the requirements and the realities of large social network datasets. Some social network analysts have even *openly* called for going beyond the traditional node-link graph paradigm (Viégas & Donath, 2004). It is clear that analysts can no longer rely solely on the traditional visualization methods—as we have done so fruitfully in the past.

At Carnegie Mellon University's Center for Computational Analysis of Social and Organizational Systems (CASOS) Lab, we tried treemaps as an alternative and have found them to be a stunningly effective. As we recount more thoroughly in a later section of this report, including treemaps in our exploratory data analysis process has been fruitful. They have provided us with an ability to scan the full information landscape of a large network, as well as providing us with focal points on underlying areas of interest. Had we not used a treemap, we very well may have missed many important features of the data that we were exploring. Other analysts may benefit from treemap technology as well, thus our sharing of this experience. We strongly encourage others to discover the benefit of using treemap technology for themselves.

In this technical report, we present the basic idea of treemap technology and propose how it might be applied broadly in social network analysis. We provide a narrative of how we have actually used treemaps in a study of a large, real-world, social network. To conclude this report, we provide possible directions for future development and application of treemap technology, specific to social network analysis.

2. Introduction to Treemap Technology

In this section, we broadly describe treemap technology; first, by explaining and showing a simple treemap image, at-a-glance, then by presenting a more complete synopsis of its

distinguishing features. Next, we provide a simple example of its use in an actual situation, and then present several variations to the basic treemap design. This section ends with a partial survey of how treemap technology is being applied in various fields, and where to find treemap software and more detailed information.

2.1 Treemaps: At a Glance

Treemap technology consists of an algorithm that results in a two-dimensional graphical representation—called a *treemap* (Johnson & Shneiderman, 1991). The algorithm arranges hierarchically-structured, categorical data, into a rectangular, mosaic image. The treemap image could be largely thought of as an amalgamation of a tree diagram and a pie chart, with additional features; it depicts both the structure and the content of the data hierarchy.

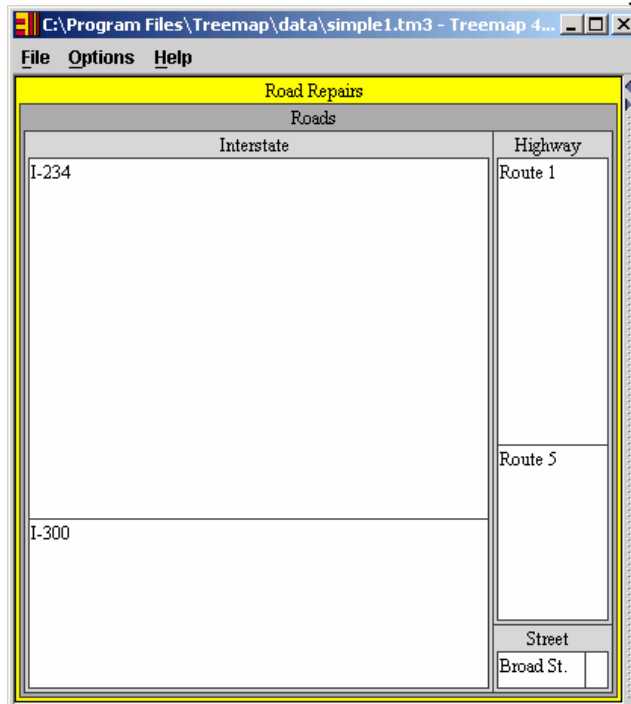


Fig. 1. A Treemap of a Road Maintenance Record (Shneiderman & Plaisant, 2003).

Figure 1 is a picture of a treemap image; this example represents a very simple dataset consisting of a two-tier hierarchy and categories, namely a road maintenance history for one year. The treemap algorithm partitions the display space into a collection of box-shaped elements, forming a mosaic. Each inner box pertains to an individual data element and is sized (the physical area the box takes in the display space) according to a quantitative value: in this case, each box represents a specific road and its size reflects the number of repairs made on the road over the past year. Each road is classified into a category—Street, Highway, or Interstate—, which is reflected on the treemap by placement of boxes. All of the elements of category highway are positioned *within* the larger highway box; accordingly, all the streets are positioned within the larger streets box, et cetera.

Quickly, the hierarchy of the roads to their categorical type can be seen; as can the relative number of repairs for each road. The most frequently repaired road can be found quickly by

looking for the largest box element. The relative number of repairs at the highest level, the road type, can also be found in the same image.

A distinctive feature of treemap technology is its powerful interactive component. The treemap is intended to be used interactively on a computer and displayed on a computer screen, rather than viewed as a static, or printed, display. This allows for effortless exploration of the data hierarchy with concurrent appraisal of the quantitative aspects of the data. (We'll demonstrate this drill-down feature using a large dataset, later.) In addition, computer screen mouseovers are often built-in to treemap software, which can provide detailed element-specific information to the user simply by placing the computer mouse over the box of the element of interest. These interactive features situate the treemap as an immensely powerful tool for visual exploration of large datasets.

The treemap was conceived since the advent of the personal computer and benefits from advances in cognitive visualization theory and widespread computerization. Treemap technology is an advancement that builds on earlier traditional ideas and has been continually refined and altered (Andrews & Heidegger, 1998) over the decade since its origin. The idea was initially conceived to simplify the difficulty of managing file space on a large computer-storage disk device (Bederson & Shneiderman, 2003). Working with long lists of filenames and space utilized, even when sorted, told only a portion of the full "story" of how a disk's valuable space is being used by multitudes of users. Computer system administrators could not easily interpret the usage from the logical perspective of the disk's directory-hierarchy, which is how computer files are typically organized. (We'll use this same application as a demonstration, later). The name, "treemap", comes from turning a computer's file directory *tree* into a planar space-filling *map* (see Shneiderman, 1998, for a first-hand account of the history and evolution of treemaps).

2.2 Characteristic Features

Johnson and Shneiderman (1991) set out four design objectives for the treemap: (a) efficient space utilization, (b) interactivity, (c) rapid comprehension, and (d) pleasing esthetics. They wanted a method for visualizing hierarchical structured information effectively in a confined display space (computer screen) that would allow for effective exploration of the underlying data. They were motivated by the lack of any methods that could meet all of the criteria in one tool. The existing methods could not meet the demands of current data exploration needs, particularly in the case of large hierarchical datasets.

A hierarchical dataset contains two kinds of information: structural and content (Johnson & Shneiderman, 1991). The *structural* component is associated with the hierarchy; usually thought of as the network of multi-level relations, as in a non-cyclic tree structure, with nodes at the terminals. The *content* pertains to content of each node, that when rolled-up its hierarchy and has meaning at the higher, grouped level(s), as well as at the individual node level.

From this perspective, there are three variables that can be represented in a treemap: (a) a quantitative value for each entity, (b) the group membership of an entity, and (c) a relative and ordered relationship of each entity within a group; respectively, these variables are represented by (a) the scaled size of the rectangle representing the entity, (b) the placement of an entity within the larger rectangle, and (c) the placement of an entity's rectangle with the parent rectangle. The placement aspect (item c) is an area that is currently being developed; so far, there seems to be inherent physical limitations to reflecting contextual meaning to the placement of entities within a parent box.

The ability for a treemap to represent a hierarchy of data within a restrained and rectangular space—such as a computer screen—makes the visual both efficient and effective. While a dataset hierarchy can be represented in the traditional form of a network tree, there is a great deal of unused space on the traditional graphic’s palette. The treemap fills the display space in its entirety and esthetically, while still showing the entire structure of the data.

We posit that the feature most significant about treemap technology is its ability to represent large datasets². Treemaps have been successfully applied to the visualization of one million items (Fekete & Plaisant, 2002). This single feature makes treemaps uniquely relevant to researchers: we will expand on this point in the section applying treemaps to social network analysis (Section 3).

Table 1. Features Comparison of Visualization Techniques

	Pie Chart	Histogram	Adjacency Matrix	Sociogram (node-link)	Treemap
Information conveyed					
Individual Nodes	No	No	Yes	Yes	Yes
Individual Ties	No	No	Yes	Yes	No
Categorical	Yes	Yes	No	No	Yes
Hierarchy	No	No	No	No	Yes
Size of Groups	Yes	Yes	No	No	Yes
Display-space utilization ¹ :	75%	50%	90%	25%	100%
Free from dataset size?	Yes	Yes	No	No	Yes
Interactivity:	Static	Static	Static	Dynamic	Dynamic

Notes: (1) Our intuitive estimate.

It is useful to compare various characteristics of treemap technology with those of traditional visualization techniques. Table 1 is a summarized, side-by-side comparison of treemaps vis-à-vis several common chart types—social network specific and not. While there are likely particular

² While it is not the explicit intention of this paper to propose methods, algorithms, or software to construct treemaps on a computer, we recognize that the practical application of treemaps rests, to a great extent, on the responsiveness of the treemap software—to compute and display the treemap with speed. While we have not stress-tested the University of Maryland (2003) implementation with extremely large datasets, we believe that extremely large datasets can be displayed quickly using sophisticated software algorithms (see Keim & Kriegel, 1996). In addition, for extremely large datasets, it is likely that a pre-processing data reduction process can be carried out—if necessary—prior to the computer-user experience, thus mitigating any sluggish display times. For an example of an underlying treemap algorithm, we suggest looking at the squarified treemap algorithm (Bruls, Huizing & Wijk, 2000).

implementations of pie charts—for example, that may, today, be interactive—, we evaluated each feature against the current atypical implementation of the technique based on our personal experiences. The table shows that treemaps convey much of the same information as other techniques, except for an ability to show individual ties (relationships) in the network, however, treemaps combine these features into one. Unlike any other technique, the treemap can make full utilization of the available display space. As mentioned earlier, the treemap technique is free from the size of the dataset and has been purposely designed for user-interactivity.

2.3 An Illustrative Example

An illustrative example of treemap technology be applied to a simple dataset provides a true sense of the technology, both its features and its effectiveness. We will walk-through an exploration of file storage on a computer disk since it is simplistic and it is likely a task common to many.

Figure 2 is an example of a treemap representation of disk space usage on a personal computer. Each rectangle represents a file on the disk. The size of each reflects the space usage of the file (or group of files). In this example, the /windows subdirectory and, recursively, its children are mapped. Because the size of each tile is meaningful, it can be readily seen that there is one file in particular that is responsible for 10-15% of the total disk usage by the /windows directory. This one file is about 45% of the size of the system32 directory, which certainly raises some question about the 28bef.msi file. Other quick observations include a large number of files in each directory—this may slow the file search speed on the computer. Also it can be seen that there are a large number of “\$” files; notice the cluster of them on the bottom right section of the map.

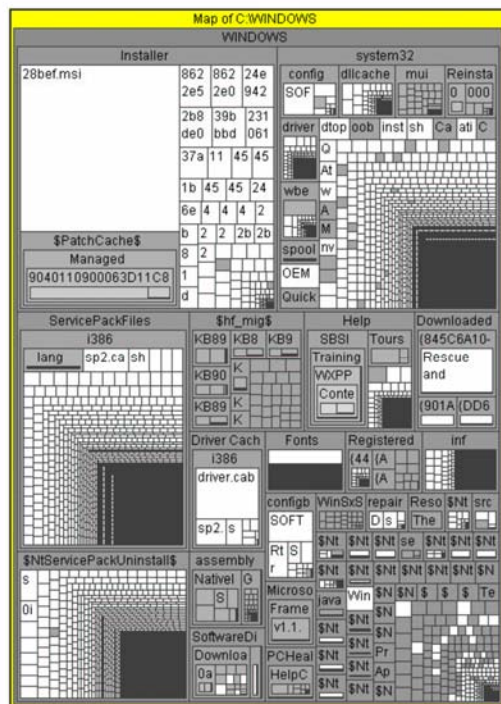


Fig. 2. A Treemap of a Computer Disk Directory: File Size and Directory Hierarchy.

The drill-down capability of the treemap is shown in Fig. 3. This is a drill-down on the /windows/Help subdirectory....you can see the relative sizes of the files and the subdirectories. In this software, a simple double-click on the screen initiates the drill down and right-click transverses upwards.

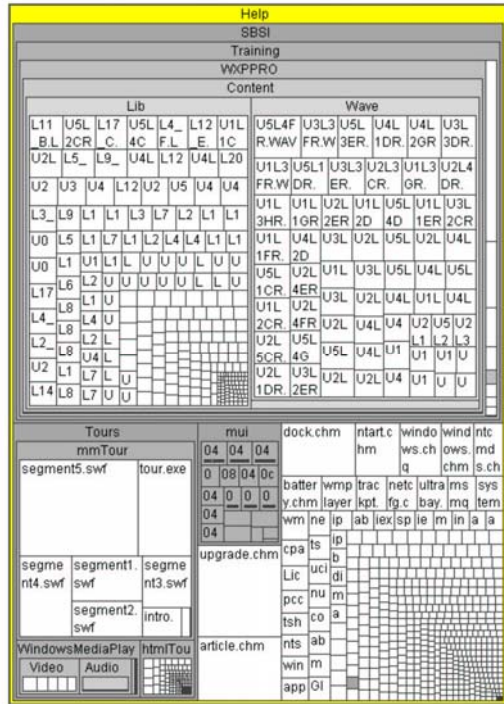


Fig. 3. Results from Drill-Down of Prior Image (Fig. 2): Treemap of the “Help” Subdirectory.

For full effect, we advocate trying the interactive treemaps in-person using your own personal computer data. Fully functional, free software is available for download on the world wide web (see Section 2.6).

2.4 Variants of the Core Treemap Technology

There are several variations in the layout of the treemap display. Different layouts are known as the slice-and-dice, the cluster, the squarified, the pivot-by-middle, the pivot-by-size layouts, and the strip; among many subtle differentiations, each layout has different visual qualities, characteristic strengths and weaknesses, and may, or may not, maintain a dimension of hierarchical order (Shneiderman & Wattenberg, 2001). A size by size comparison of randomly generated slice-and-dice, squarified, strip, and pivot-by-size treemap layouts (Bederson, Shneiderman & Watterberg, 2002; Wattenberg & Bederson, 2005) demonstrates the subtle differences between the types and the different impact each type has visually on interactive users.

Recently, the treemap algorithms have been extended, beyond displaying static, but still *manually* interactive, views of the data, to being capable of *automating* the transition from one representation of the data to the next (Ghoniem & Fekete, 2001), which is particularly useful in real-time data analysis. Treemaps have also been combined with Compound-Fisheye Views—a multi-level tree-graph cluster visualization technique—to display large network graphs (Abello, Kobourov & Yusufov, 2004), albeit in a different style and context than we discuss in this technical report.

2.5 Application in Various Fields

The treemap, along with recent derivations, has been applied in a broad range of applications in seemingly divergent disciplines. In financial services, most notably in the stock market, they are used to explore the real-time stock-market performance of the entire stock universe, at a glance. The SmartMoney (2005) Map of the Market website is a freely-accessible implementation of treemaps for this purpose. Other applications include sports analysis (Turo, 1994, 2003), organizing photographs (Bederson, 2001), global news tracking (Newsmag, 2005), among many others.

2.6 Software and More Information

Treemap software and software libraries are freely available from several sources. We suggest taking a first look at the SmartMoney Map of the Market site (SmartMoney, 2005) to see an excellent implementation of treemaps (and the derivation heatmaps) that has become popular in the investment community. Next a visit to the treemap webpage at the University of Maryland (2005) for fully functional treemap software that can display your own data, as well as the underlying software libraries that allow you to expend your own custom software to include treemap technology.

3. Application to Social Network Analysis

While traditional, time-honored network visualization tools, e.g., node-link diagrams, bar charts, tree diagrams, etc, remain indispensable, treemaps have direct benefit in specific areas of social network analysis—we emphasize, however, that treemap technology ought to be considered an *adjunct* to the traditional tools and *not* thought of as a replacement. In particular, treemaps likely are a best-fit if used during exploratory data analysis, particularly when examining clustering models—*cluster models* are settings where nodes are partitioned into groups based on specific criteria then the network studied across or within the groups.

The software indispensable for creating node-link diagrams automatically places the nodes using a specific algorithm to provide some information according to the placement of the nodes, either relative to the center of the screen, and / or relative to one another. The nodes can be placed, using multidimensional scaling, or according to any of several other important algorithms, even random placement, if desired. Regardless, of which placement algorithm is applied, the use of node-link diagrams have an inherent problem with space utilization. In the case of large-sized networks, the analysts are likely presented with a useless visual blob. To date, the widely available software has little capability for grouping exploration in an interactive and visual fashion. Overall, treemaps are well suited for overcoming these limitations of current graph-link diagrams. The treemap's design characteristics: efficiency of computer screen display, the ability to effectively present large networks, and the effectiveness of its interactive capabilities, all make treemaps well suited for exploration of social network groups.

When relative size of a group or subgroup(s) are of interest, such as in analyzing hierarchical networks, or cellular networks.

In this section, first we identify several characteristics of treemaps that make their use beneficial to social network analysis and, then, as an example of their actual use, we describe how we have actually applied treemaps while exploring a large real-world social network dataset.

3.1 Treemaps in the Social Network Setting

Among the many beneficial traits of treemap technology, five characteristics are directly relatable to social network analysis; treemaps: (a) offer a visual representation of quantitative features of a clustering model, (b) show the hierarchical aspects of a clustering model, (c) provide for interactive exploration of the hierarchy of a clustering model, (d) make full use of available display space, and (e) are unaffected by the size of the underlying network. On their own, these characteristics are not unique, but a treemap uniquely combines all of these benefits into a single tool, thus its attractiveness to analysts.

The first related aspect of the treemap is that the mosaic layout of treemap data allows for visual examination of quantitative features of a clustering model that most traditional methods do not provide. Its layout facilitates visual comparisons of subgroups, both between the set of groups and for each subgroup in relation to the whole. The relative size of each tile (the spatial area of each rectangle within the entire treemap display space) is readily apparent on the treemap and represents a quantitative attribute of the group; often this attribute is based on the group-membership count statistic, but it is not restricted to counts. The value of a social network measure pertaining to the group can be assigned to the tile's area value. The constraint, however, is that the values for all tiles must total unity; it may be necessary to normalize the data to accomplish the unity requirement. For example, the area of the tile can be determined by the percentage of relationship ties the members of the group is adjacent to in, relative to the total number of ties in the network.

In many ways, the visual information presented by the treemap, in a static view, is consistent with that of a typical bar or stacked-bar chart, but instead of the data tiles being spread out, or stacked atop one another, on an x-y axis, the treemap neatly places the assortment of tiles within a single bounding rectangle. The area of each tile in the mosaic is sized according to a quantitative value with its physical area calculated relative to the whole display space and 2D dimensions based on a selected algorithm to control the aspect ratio, for example, squarified (see prior section for a discussion on the types of displays). As designed, the treemap displays the quantitative information associated with each group, but simultaneously, it also can display important structural information, which is discussed next.

A second characteristic relatable to social network analysis is the treemap's ability to construct a complete visualization of hierarchically structured data. The 2-D display rectangle, as well as each sub-tile, can be recursively sub divided further into sub tiles, ad infinitum. Users can choose to display any number of the levels simultaneously, or alternatively hide levels in the visualization. In this respect, the treemap displays nearly the same information as a traditional tree diagram, organization chart, or a dendrogram—a *dendrogram* is a tree-like graphic presented to represent the link results from discrete similarity-based hierarchical clustering algorithms, such as CONCOR. We consider a treemap as a practical and more powerful alternative to the traditional dendrogram graphic, because of the treemaps simultaneous visualization of hierarchical structure *and* relevant quantitative information.

A third characteristic of treemap technology beneficial to social network analysis is its inbuilt design for being interactive, principally operated on a computerized platform with a display screen. The interactive control over its display of information, with real-time feedback, is essential to the *exploration* of complex data (Johnson & Shneiderman, 1991). Most data visualization techniques were invented prior to widely-available computer resources, many even created in the pre-computerized world. Treemaps were conceived as a result of computerization, and certainly have benefits from this more modern scenario. The interactivity feature gives the analyst the ability to transverse the tree and be presented a view of the categorical data at each level.

The next beneficial characteristic of treemap technology on our list arises from its aim to make full use of the available display space. Most desirable, today, is the ability to pack as much information into as small a space as possible, especially displays on a computer screen. The treemap provides the hierarchical categorical information in the 2D display in a 100% space-filling manner. We evaluate this feature vis-à-vis the display space inefficiency of the traditional x-x-y axis bar chart, tree diagram, or the atypical node-link diagram. The space-filling characteristic is especially useful in the confines of a computer screen and under the trend towards data dashboards. We can even posit that the treemap is *more than* 100% efficient in its display space utilization. When interactive mouse-over windows are programmed into the treemap graphic, practically any information can be presented in these small display screen overlays.

The last on our list of characteristics of treemaps that are directly relatable to social network analysis is the technology's ability to work with large datasets. Treemap technology scales to a theoretically unlimited number of data points and has been shown to work well with nearly one million items on a single display (Fekete & Plaisant, 2002). With the size of networks under study increasing, analysts no longer can glean useful information from traditional node-link diagrams. Using the digital computer screens, tiles may be displayed as small as a single screen pixel, but with the interactive capabilities built into treemap technology, the user can zoom-in on the tile(s) of interest and expand the tile to the full size of the entire display space if desired.

While these five characteristics of treemaps are not unique to treemaps, only treemaps combine all of these characteristics into a single tool. These characteristics are particular beneficial when exploring subgroups in a network.

3.2 Examining Subgroups

A frequent question asked about a social network is, “what are the cohesive node-subgroups?” (Wasserman & Faust, 1994). Which is often followed by, “and, how do the groups compare?” To answer these questions, analysts apply their judgment or formal clustering techniques to partition nodes into groups. They then examine the results; we call these results the *clustering model*, which re-represents the network as separated collections of nodes, each node being assigned membership in a distinct subgroup.

Formation of the subgroups can be accomplished either by a manual approach of methodically or ad-hoc manually exploring the possibilities in a hit-or-miss manner, or by way of using computational or statistical methods. Regardless of how the nodes are clustered into groups, the analyst will also explore the results.

To examine the clustering results and subgroups, traditionally analysts rely on visual graphics for much of the ad-hoc exploration. Conspicuously, Moreno (1932) used a visual (a

node-link diagram) for his analysis in the very first, social network study. Still today, most often, a basic node-link diagram is relied upon for this discovery step.

When examining the results of the clustering process, the analyst will likely look over the membership of the subgroups and make comparisons across and between the groups. While treemaps can be an aid in the *manual* clustering process, they are not designed to be of any use in the statistical approach. However, treemaps are an indispensable aid in the post-clustering inspection step. We will describe these two applications in detail below.

Entities in a treemaps may be one of either a single actor, a.k.a., node, or a meta-node (a set of nodes purposefully grouped together and represented and treated as a single node in a social network)

First, to aid in the manual clustering method, treemaps are supplemental by their ability to display both hierarchal and categorical information simultaneously, and interactively. Analysts stipulate node clusters by specifying categories based on the values of the specific attribute, By setting the hierarchy determination criteria to a simple categorical attributes of actors in the clustering process, the treemap can interactively show the results in a hierarchical form which allows a instant visual for the analyst to evaluate the criteria.

A simple clustering could be formed and examined based on descriptive categorical values, such as by gender, ages, nationality, job rank, and so on. More social network related categories could be applied such as the number of an actor's incoming, outgoing, or bidirectional ties. More involved categories can be used by using social network measures for an actor and binning the value into a number of bins and clustering according to bin membership.

Each of these options, many based on the data actually available, can be easily created using simple treemap software—without using much more mathematics than counting. Using treemaps, the analyst can visually examine, perhaps almost instantly, the results of applying the attribute-based clustering methodology to the data. A basic observation of the number and relative size of the tiles in the treemap can provide a visual clue to the results of the clustering policy.

Actors are segregated, or clustered, into distinct groups and analyzed for uncharted similarities and relative size of the group. While treemaps, per se, are not designed to, nor will they identify the membership sets of the groups—we'll leave that to other tried and true techniques—, treemaps will facilitate the exploration of the groups and their characteristics.

Second, treemaps are highly effective in the examination of subgroups after the clustering has been performed. In addition to examining subgroups created by the manual process above, treemaps can be useful in examining the clustering created by more hidden mathematical and statistical methods.

The analyst can transverse a hierarchical tree of clusters which are a product of the statically methods, and are well-suited for examining the results from procedures such as hierarchical clustering algorithms. The treemap allows for an interactive transversal of the hierarchy tree, showing the corresponding groups and their relative areas at each level.

One successful visual to explore subgroups in a network, is the dendrogram graphic. The dendrogram is typically used to display the results of a hierarchical clustering algorithm. CONCOR is another powerful grouping algorithm that also places actors into distinct subgroups in a hierarchical fashion and can also be displayed via a dendrogram.

We hold the view that treemaps are of most value as an interactive visualization tool. A static picture of a treemap provides a great deal of information, for the most part, the same

information can be conveyed visually in traditional graphical techniques, such as bar charts, line graphs, etc.

4. Demonstration of Exploring a Large Social Network

In this section, we recount how we conducted an exploratory data analysis of an actual social network dataset. We describe only pertinent steps for this overall discussion; first looking at a traditional node-link diagram, then using the same data, we explored using a series of treemaps. The research was conducted at our lab for the purpose of academic inquiry into an organization's email communication network, thus these activities are not expressly intended to systematically trial or evaluate treemap technology, per se.

The assortment of steps we describe are not all-inclusive of how treemaps can be used, but this does provide first-hand experiences that serve as a demonstration of their effectiveness. We used treemaps in the exploratory data analysis phase of our study—herein, we'll refer only to our visual exploration tasks. The quantitative perspective for our analysis involves only the number of ties an actor has—either in or out ties. All examples in this narrative pertain to only this quantitative measure.

We were investigating the Enron email corpus (Diesner, Frantz & Carley, 2005). The data we explored represents hundreds of thousands emails which were sent and received by Enron employees between 1999 and 2002. We examined an enriched version of the full dataset that also contained full names mapped to email address(es) and job titles mapped to the full names.

4.1 Using Traditional Diagram

Habitually, our first step in the visual exploration phase was to pursue the conventional path of inspecting the social network with a traditional node-link diagram. We used the powerful ORA social network statistics and visualization software as we commonly do for all our research. Of course, there are many other software programs that could have been used for this first step, as we simply wanted to visualize the network.

Figure 4 is the node-link visualization from ORA. From this picture, it is clear that the social network is indeed complex, with very many nodes and corresponding ties. The actor nodes appear in the foreground as colored dots with a background looking like a mess of lines. We can make few, if any, useful observations from this diagram.

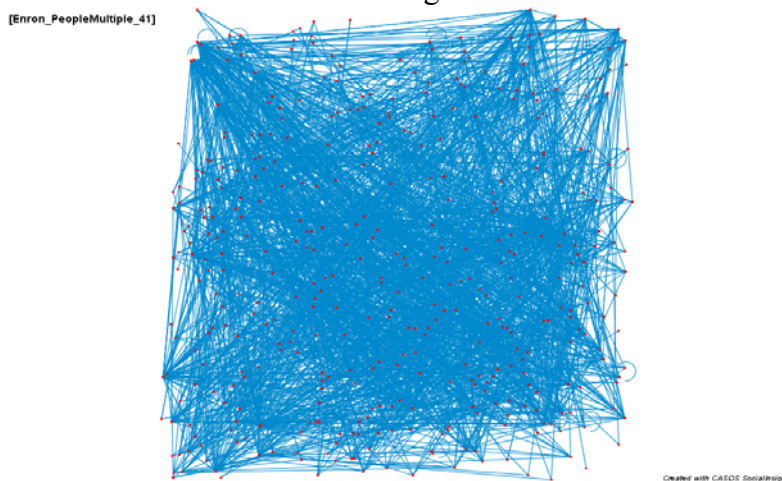


Fig. 4. Node-Link Diagram of a Large Social Network Dataset: The Enron Email Corpus.

Interactively, we repetitively filtered the visual on the number of ties and removed isolates and pendent nodes each cycle. There we dozens of network pictures to look at; some of the pictures were difficult to really glean and useful information. Figure 4 is an example of one of these k-core diagrams.

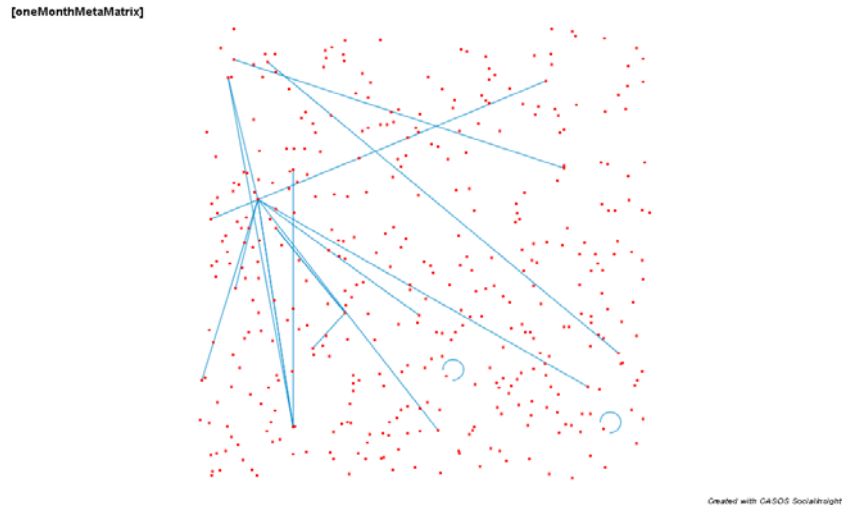


Fig. 5. Drill-Down of a Node-Link Diagram: The Enron Email Corpus.

Next we took a look at a histogram of the degree for the nodes, as shown in Figure 6. Here we located the high-degree nodes for a single month but were not easily able to assess this one month in relation to other months. We would have needed to print each of the 48 months and compare side by side on a very large table. Alternatively we could have created a side by side bar chart with thousands of vertical bars, which would have proved ineffective.

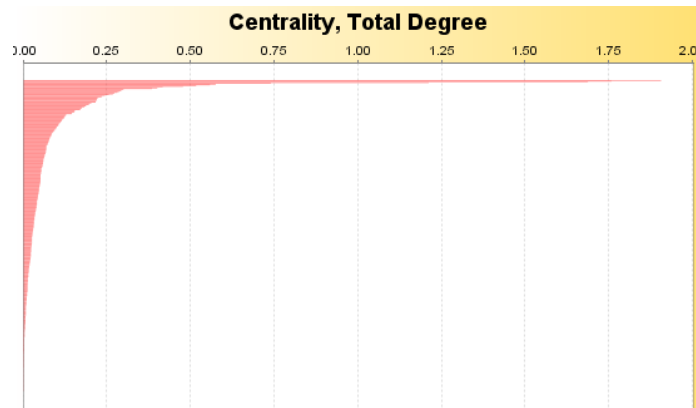


Fig. 6. Traditional Histogram of Node Degrees: The Enron Email Corpus.

We surmise that our experience is atypical to what many social network analysts experience in the exploratory data analysis phase; of course, there is likely a wide variation depending on the experience and inclinations of any specific analyst. Our analysis from this point forward, we will not provide here as it is unrelated to the topic of this report. Instead, we will move to using treemaps technology as a continuation of the exploratory data analysis phase.

4.2 Using Treemaps

Next, we loaded the network data into Treemap software (University of Maryland, 2003). To accomplish this, it was necessary to transform the network data into XML format; a format specific to the particular treemap software requirements. We did have to create several versions of the treemap data files according to the hierarchies that we pre-selected that may be of interest for this dataset. This multiple pre-formatting step is a function of the particular software we used and not of treemap technology, per se. Other treemap generation software packages may not require this added data formatting step for each of the different views of the network hierarchy.

We started the exploration with a treemap of the network wanting to get a sense of the network according to the job titles and persons within the various titles. Figure 7 shows the first treemap. From this treemap we can see that Vice Presidents, Employees, and Directors all have about the same number of ties (The area of the corresponding titles are about the same size). Further, we can see that Employee Dasovich has the greatest number of ties in the entire network. We can easily see that two of the four Enron presidents have equal values and are much larger than the other two. We can also see that the lawyers communicated nearly as much as the Managers and Presidents combined, even though there are more managers than lawyers.

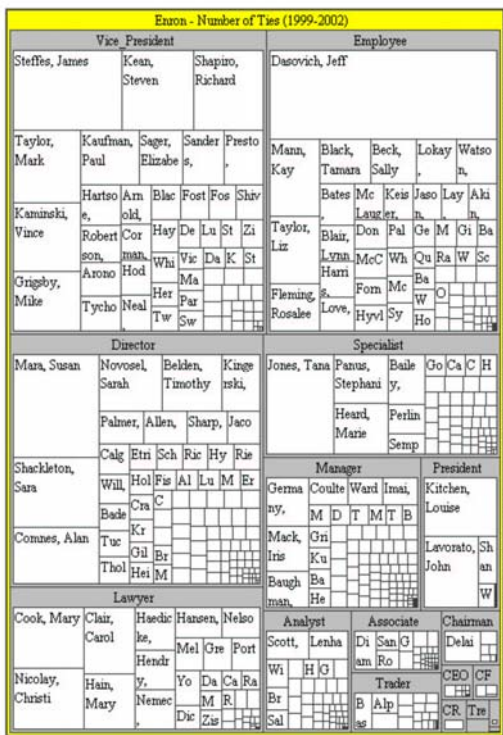


Fig. 7. Treemap: Actor Node Degrees, by Title

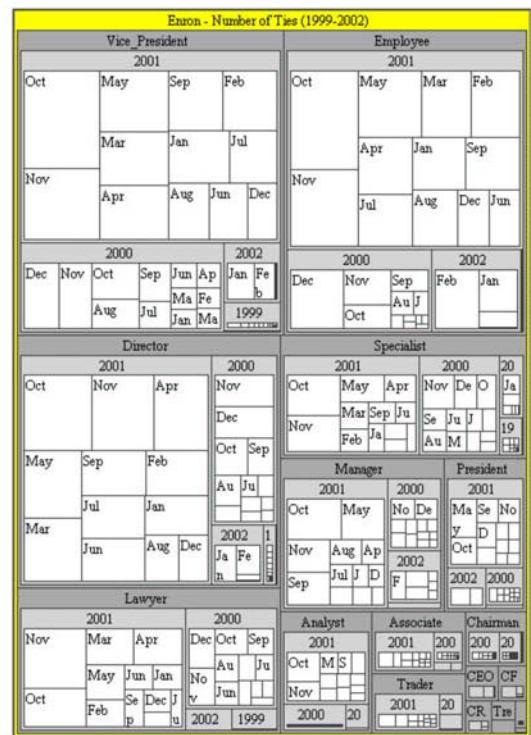


Fig. 8. Treemap: Actor Node Degrees, by Title, Year, & Month

Next, we wanted to get a sense of the temporal aspect of the data. We looked at a treemap showing the hierarchy based on title, then date. Figure 8 shows this treemap. Quickly we saw that the largest amount of communications was in October 2001 by the Vice Presidents. We also noted that October was the largest for practically all of the job titles.

To see a more direct view of the temporal aspects, we looked at a treemap with the hierarchy by year, month and individual. Figure 9 shows this treemap. Quickly we see that 2001 was by far has the most ties and that October and November combined account for about 1/3 of all 2001 communications and perhaps about 1/4 of the communications for the entire dataset. We see that 1999 is a mere fraction of the activity in the dataset.

To get a more focused view of 2001, we simply drilled down into year 2001 by clicking the “2001” (thanks to the capabilities of the software). Instantly we had view of only this one year and we could drill down further to a particular month, a specific day, ultimately even to a specific email. Figure 10 shows the results of drilling down to 2001. We have a more detailed picture of 2001 and can see that Dasovich was superseded by Mary Cook in November. Perhaps Dasovich took vacation that month, or maybe Mary Cook was an assistant of Dasovich and proxied his email communication during the period. While the actual answers to these questions require more precise investigation, the treemap exploration does quickly provide areas of interest.

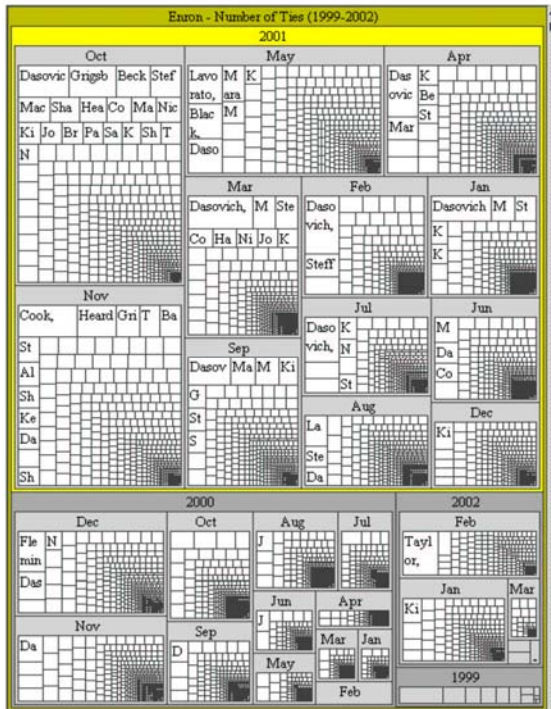


Fig. 9. Treemap: Node Degrees, by Year & Month

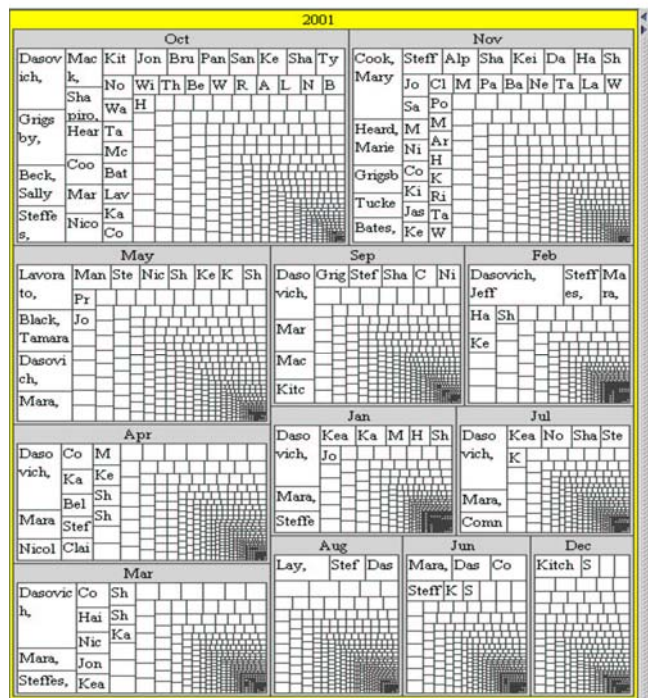


Fig. 10. Treemap: 2001 Node Degrees, by Month

Figure 11 is a further drill down of the same treemap, this to the next level of showing only a single month, October 2001. With this view we can see more names clearly and see that there perhaps an exponential distribution of the ties.

Figure 12 is October 2001, again, but this time looking at the job title, then individual name hierarchy. Here we see that Vice Presidents and Employees had about 45% of the communications during the month and that the directors communicated about 20% of the total.

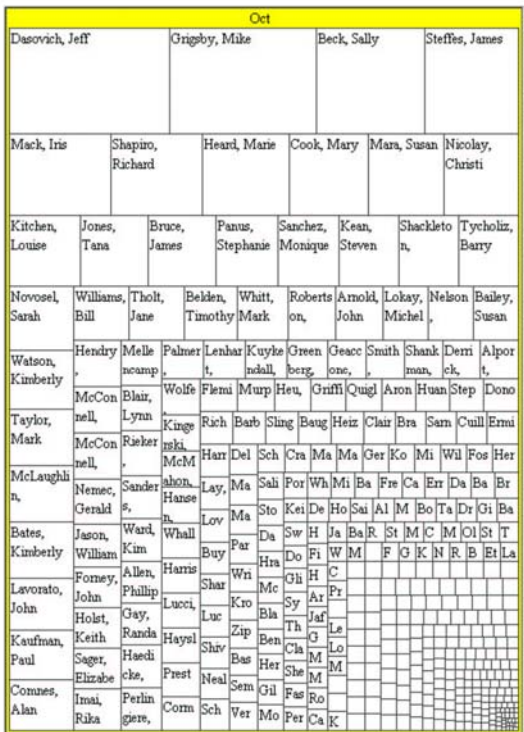


Fig. 11. Treemap: Oct. 2001 Node Degrees, by Name

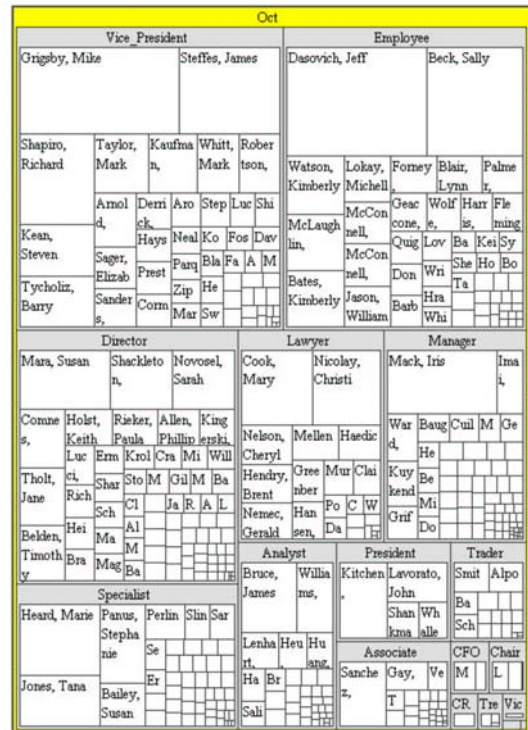


Fig. 12. Treemap: Oct 2001 Node Degrees, by Title, & Name

4.3 The Overall Experience

It was quickly very evident that using treemaps added an immense value to the exploratory data analysis phase of our research. Particular when put side by side with one of the traditional visualization techniques, the immense improvement of treemaps shown through very clearly.

There were only a few questions sparked from exploring the node-link diagram, while exploring the dataset using treemaps, generated dozens of interesting observations and follow-up questions. The perspectives were introduced to from the treemaps *popped* out at us, while the node-link diagrams required us to search around for interesting issues and only set up for broad questions. The treemaps gave us a view that provided not only broad questions, but very targeted ones as well.

5. Conclusions and Next Steps

Our personal experience from using treemaps for exploring large social networks has been positive and analytically beneficial. We found that applying treemaps to social network analysis,—during exploratory data analysis—is effective. To provide treemap technology to others in the broader social network research community, early development is underway to implement treemaps in our Lab’s social network statistical and visualization software: the Organizational Risk Analyzer (ORA). See <http://www.casos.cs.cmu.edu/projects/ora> for documentation and free downloads.

We also believe that a derivative of treemaps, the heatmap, may hold still more promise than treemaps; heatmaps appear to be an even richer exploration tool for social network analysis. The heatmap technology expands on the treemap by utilizing color and symbols to increase the amount of information visible, at a glance, to the user. We foresee trying heatmaps in the future, and expect that, like treemaps, heatmaps will prove to be a valuable tool to our analysis activities.

In conclusion, when working with treemaps, it quickly becomes evident that for specific analytic purposes, the treemap is an effective tool and is an instantly recognizable advancement in visualization technology that has clear-cut benefits to social network analysis. We encourage others in the social network analysis community to try treemap technology and render a *positive* opinion for themselves.

6. References

- Abello, J., Kobourov, S. G., & Yusufov, R. (2004). Visualizing Large Graphs with Compound-Fisheye Views and Treemaps. *Proc. Graph Drawing 2004*, pp. 431-442, New York City, New York, USA, September 29 - October 2, 2004. Springer-Verlag.
- Andrews, K., & Heidegger, H. (1998). *Information slices: Visualizing and Exploring large hierarchies using cascading, semi-circular discs*. IEEE Symposium on Information Visualization (infoVis '98), Research Triangle Park, North Carolina, October.
- Bederson, B. B. (2001) *PhotoMesa: A zoomable image browser using quantum treemaps and bubblemaps*. UIST 2001, ACM Symposium on User Interface Software and Technology, CHI Letters, 3(2), 71-80.
- Bederson, B. B., & Shneiderman, B. (2003). Understanding Hierarchical Data, Chapter 6 Introduction. In B. B. Bederson & B. Shneiderman (Eds.), *The craft of information visualization: Readings and reflections*, pp. 229-233. San Francisco: Morgan Kaufman.
- Bederson, B. B., Shneiderman, B., & Wattenberg, M. (2002). Ordered and Quantum treemaps: Making effective use of 2D space to display hierarchies. *ACM Transactions on Graphics*, 21(4), 833-854.
- Bruls, M., Huizing, K., & Wijk, J. J. v. (2000). Squarified Treemaps. Proceedings of VisSym '00 (May 29-31, Amsterdam, The Netherlands), Springer Wien New York, 33-42.
- Diesner, J., Frantz, T., & Carley, K. (2005). Communication networks from the Enron email corpus: “It’s always about the people. Enron is no different.” *Computational & Mathematical Organization Theory*, 11, 201-228.

- Fekete, J.-D., & Plaisant, C. (2002). Interactive information visualization of a million items. *Proceedings of IEEE Symposium on Information Visualization 2002 (infoVis 2002)*, Boston, October.
- Freeman, L. C. (2000a). Visualizing social groups. American Statistical Association, 1999 *Proceedings of the Section on Statistical Graphics*, 2000, 47-54.
- Freeman, L. C. (2000b). Visualizing social networks. *Journal of Social Structure*, 1(1).
- Freeman, L. C. (2005). Graphic techniques for exploring social network data. In Peter J. Carrington, John Scott and Stanley Wasserman (eds.), *Models and Methods in Social Network Analysis*. Cambridge: Cambridge University Press.
- Ghoniem, M., & Fekete, J.-D. (2001). Animating treemaps. *Proc. of 18th HCIL Symposium - Workshop on Treemap Implementations and Applications* (May 31, University of Maryland, College Park, Maryland, USA).
- Johnson, B., & Shneiderman, B. (1991) Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the 2nd International IEEE Visualization Conference*, pp. 284-291, San Diego, CA.
- Keim, D. A., & Kriegel, H-P. (1996). Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6).
- Newsmap. (2005). www.marumushi.com/apps/newsmap/
- Shneiderman, B., & Plaisant, C. (2003). *Treemap 4.1 software*. University of Maryland. www.cs.umd.edu/hcil/treemap
- Shneiderman, B., & Wattenberg, M. (2001). Ordered treemap layouts. *Proceedings of the IEEE Symposium on Information Visualization 2001*, 73078. Los Alamitos, CA: IEEE Press.
- SmartMoney. (2005). *Map of the Market*. www.smartmoney.com/marketmap/
- Tukey, J. W. (1977). *Exploratory data analysis*. Philippines: Addison-Wesley.
- Turo, D. (1994). Hierarchical visualization with Treemaps: Making sense of pro basketball data. *ACM CHI'94 Conference, Boston, MA, April 24-28*.
- Turo, D. (2003). Hierarchical visualization with treemaps: Making sense of pro basketball data. In B. B. Bederson & B. Shneiderman (Eds.), *The craft of information visualization: Readings and reflections*, pp. 237-238. San Francisco: Morgan Kaufman.
- University of Maryland. (2003). *Treemap*. Software available at www.cs.umd.edu/hcil/treemap
- Viégas, F. B., & Donath, J. (2004). *Social network visualization: Can we go beyond the graph?* Workshop on Social Networks, CSCW'04, November 6-10, 2004. Chicago, IL, USA.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, MA: Cambridge University Press.
- Wattenberg, M., & Bederson, B. (2005). Dynamic treemap layout comparison. Retrieved 1 September 2005 from www.cs.umd.edu/hcil/treemap-history/java_algorithms/LayoutApplet.html