Information Content Measures of Visual Displays

Julie Yang-Peláez and Woodie C. Flowers
Dept. of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139
jayang@alum.mit.edu, flowers@mit.edu

Abstract

With an increase in the number of different visualization techniques, it becomes necessary to develop a measure for evaluating the effectiveness of visualizations. Metrics to evaluate visual displays were developed based on measures of information content developed by Shannon and used in communication theory. These measures of information content can be used to quantify the relative effectiveness of displays.

1. Introduction

Given that a sender wants to communicate a set of data to a receiver, the question is what is the most effective representation of the data. The encoding of data into a visual display can take on several possible forms, but the primary function of visualization is to display information to facilitate understanding. Although there has been much written about guidelines for good graphic design practices [1-5], little is known about how to objectively choose the better display overall.

Designers could benefit from metrics for evaluating and choosing visual displays of information. Metrics have been proposed by Brath for three-dimensional visualizations [6]. Card and Mackinlay facilitate comparisons of visualizations by categorizing the visual data types present in displays and presenting this information in morphological tables [7]. In this paper, we introduce metrics using information content measures based on Shannon's mathematical communication theory or information theory. Quantifying the information content of displays is a new approach towards the goal of measuring the effectiveness of visualizations.

2. A representational framework

Before beginning our discussion on information content measures, we must first establish a framework for the visual representations in order to have a common understanding.

First, in the communication of visual information between sender and receiver, visualizations can be analyzed at different levels: syntactic, semantic and pragmatic. The syntactic information of a visualization can be defined as the marks, lines, regions and their organization in the display. However, a syntactic analysis does not involve an interpretation of what the display represents. The semantic information associated with a visual display is the meaning imparted by the marks and their configurations. It is the receiver that perceives, processes, and makes a semantic reading of the display that can change as different visual representations convey information at differing degrees of effectiveness.

The pragmatic level of visual communication deals with the actual communication situation and the value or usefulness of the display above and beyond the direct semantic interpretation. This level accounts for the individuality of the sender and/or receiver and their respective experiences or schema, as well also the limitations of the communication medium.

3. Information content of visual displays

The theory of information introduced by Shannon deals with the transmission of signals through a communication channel [8]. In his treatment, information is dealt with only on a syntactic level. No weight is given to the semantics of the transmitted information. Although Shannon explicitly states that his theory does not in any way take into account the semantic aspects of information, considerable confusion and misinterpretation has resulted because of the loaded meaning attached to the word *information*. However, within these limitations, the Shannon theory can be applied in the context of visualization.

The practical attempt to determine the information content for a visual display requires a strict definition of information content. It is important to understand what it is that is being measured. For visual displays, we have identified four types of information content measures. These are the amount of information spanned by the data; the amount of information spanned by a display (or the capacity of a display); the amount of information in a particular data display; and the amount of topological information content.

3.1. The information content spanned by the data

For each set of data, the dimensions are the basic elements of a display. For example, consider Table 1, a

0-7695-0804-9/2000 $10.00 © 2000 IEEE

simple product development dataset with five dimensions: task, date, duration, resource, and status.

Table 1. The 5 dimensions of a project data set.

Task	Date	Duration	Resource	Status
Define customer needs	1/1	5 days	A	complete
Concept generation	1/6	2 days	B	complete
Design product	1/6	1 wk	C	complete
Produce prototype	1/11	2 wks	A	not complete
Test prototype	1/16	10 days	B	not complete

Each dimension contributes to the total information content of the data. The amount of information spanned by the data is the sum of the information contents for each of the dimensions. The information content of each dimension of the data set can be calculated as

$$I = \log_2\left(\frac{range}{precision}\right) \tag{1}$$

The information content associated with a choice of 2 states, as in the example of task status is

$$I_{status} = \log_2 2 = 1\,\text{bit} \tag{2}$$

The information content associated with a choice of 3 states, as in the example of 3 resources

$$I_{resource} = \log_2 3 = 1.6\,\text{bits} \tag{3}$$

The information content associated with a ratio or interval variable is dependent on the range and precision of the scale. For the task duration dimension, the range is 1 to 14 days with the precision equal to 1 day. This gives 14 increments with an information content of

$$I_{duration} = \log_2 14 = 3.8\,\text{bits} \tag{4}$$

The total information content of a data set is the sum of the individual components given the independence of the components. Thus, the total information content for the data in Table 1 is 13.4 bits and is the total amount of information that should be displayed in one graph.

3.2. The information content of a data display

If the size of the data set or number of dimensions is too large, it may not be possible to represent all of the data in one display. It may be necessary to split the data into separate displays resulting in a higher overall information content for the multiple displays. For example, it is difficult to create a single display that can display an entire database. Therefore, many smaller subsets of the database along with their key attributes would be displayed instead.

Figure 1 shows a visual display commonly known as a Gantt chart for the product development data set. The total information content for this particular display is 13.4 bits. The contributions from the individual dimensions are shown on the graph.

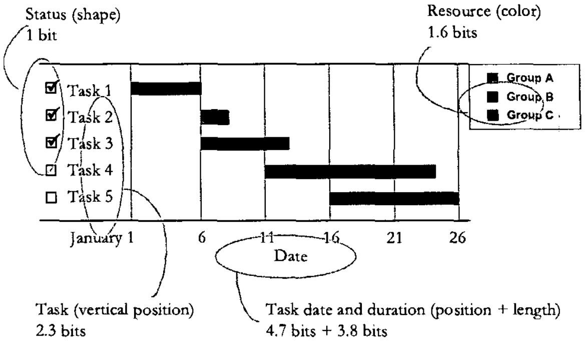
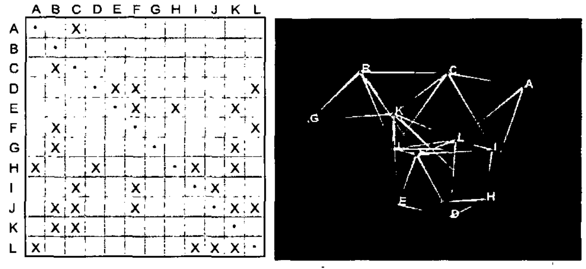


Figure 1. A Gantt chart showing data of Table 1.

In general, to create an effective display of information, the information contents of the data and the display should be equal. In this particular example, the Gantt chart effectively represents the data.

3.3. Information capacity of a display

It is useful to measure the amount of information that can be contained in a display. This measure would represent the capacity of a display. For example, suppose we wanted to represent relational information such as the task dependencies and operational flows in a product development process. These task-dependencies can be represented in a 12 x 12 matrix as seen in Figure 2a and is known as a design structure matrix or DSM. In a different representation, the information could be shown in a nodal graph as in Figure 2b.



	A	B	C	D	E	F	G	H	I	J	K	L
A	•		X									
B		•										
C		X	•									
D				•	X	X						X
E					•	X		X			X	
F		X				•						X
G		X					•				X	
H	X			X				•	X		X	
I			X			X			•	X		
J		X	X			X				•	X	X
K		X	X								•	
L	X								X	X	X	•

(a) (b)

Figure 2. a) A DSM of 12 tasks (A through L) showing task dependencies and b) a nodal graph representation.

For the general case of an N x N matrix, there are N^2-N possible locations for an X or a blank to represent the existence of a task dependency. Each location contributes 1 bit of information giving a total of

$$I = N^2 - N \text{ bits} \tag{5}$$

For the 5 x 5 matrix shown in Figure 3a, there are 20 possible locations (the boxes on the diagonal are not used) for a total of 20 bits for the matrix display.

The information capacity of an XY planar display is dependent on resolution. The grid of Figure 3b delineates the space to provide 400 possible locations for node placements for a total of 400 bits. The capacity of a display increases dramatically for higher resolutions.

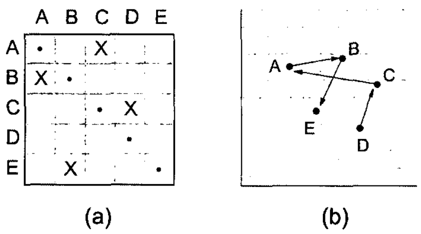


Figure 3. a) Matrix representation and b) planar nodal representation for showing relational information.

3.4. Examples

The information content associated with putting m marks on a display is

$$I = \log_2\left(\frac{p!}{m!(p-m)!}\right) \tag{6}$$

where m = number of marks, and p = number of possible locations. Therefore, for 5 tasks and 4 relations, the information associated with the matrix in Figure 3a is

$$I_{marks} = \log_2\left(\frac{20!}{4!16!}\right) = 12.2 \text{ bits} \tag{7}$$

The XY-plane representation (Figure 3b) requires more information. The information content for this display can be calculated by the sum of the information associated with displaying the nodes and the information associated with displaying the links between the nodes.

$$I_{total} = I_{nodes} + I_{links} \tag{8}$$

In the XY-plane, there are X x Y possible locations for N nodes. Therefore, the information associated with placing N nodes on the display is

$$I_{nodes} = N \log_2(XY) \text{ bits} \tag{9}$$

$$I_{nodes} = 5 \log_2(400) = 43.2 \text{ bits} \tag{10}$$

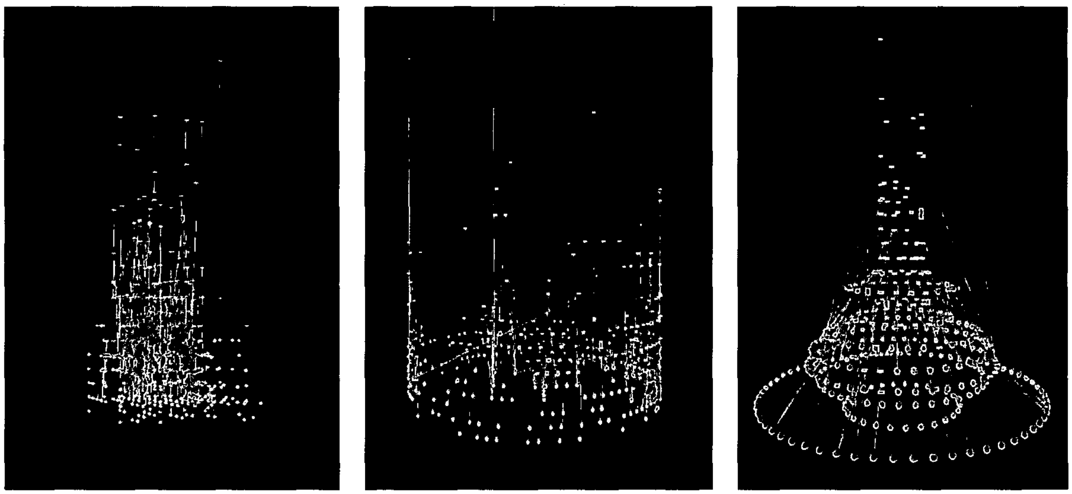
For N nodes, the number of possible directed links between the nodes is N^2-N. The information associated with the links is the same as Equation 6.

$$I_{links} = \log_2\left(\frac{20!}{4!16!}\right) = 12.2 \text{ bits} \tag{11}$$

$$I_{total} = I_{nodes} + I_{links} = 55.4 \text{ bits} \tag{12}$$

Therefore, matrix visualization requires less information (12.2 bits) for representing this relational information as compared to the node and directed link visualization (55.4 bits). The amount of information increases even more for three-dimensional visualizations.

The information content measure of a display can be used to explain the differences in the 3D visual representations of a document database. The three different representations are shown in Figure 4.



X encodes nothing, high I
Y encodes nothing, high I
Z encodes year, I_z

(a) planar

R encodes nothing, fixed, I = 0
θ encodes nothing, high I
Z encodes year, I_z

(b) cylindrical

R encodes frequency, lower I
θ encodes nothing, lower I
Z encodes year, I_z

(c) variable radius

Figure 4. Different citation graph representations.

In this particular visualization, nodes represent documents and the links between the documents represent citation references. The Z-axis encodes the year of the document for all three representations, all document nodes lie in the same plane for their corresponding year. Therefore, the Z-axis contributes the same amount of information, I_z, to each display. In Figure 4a, the nodes are arranged in XYZ space. In the planar view, the X and Y axes encode nothing. In other words, the syntactic representations of the nodes have no correspondence to semantic meaning. Therefore, both the X and Y axes contribute a relatively high amount of information.

In Figure 4b, the cylindrical representation results from a fixed radius and a variable angle. Since the radius is fixed, there is no information associated with the radius. The angle, θ, encodes no semantic meaning and therefore also contributes a high information content.

Finally, in Figure 4c, the radius in this representation is used to encode the number of documents published in a particular year. Fewer documents for a given year result in a smaller ring radius. Also, although the angle in this representation encodes nothing, the angle in this representation contributes less information because the size of the rings can be smaller resulting in larger angle increments.

3.5. Topological Information Content

The topological information content of relational information can be determined with the use of an approach described by Rashevsky in the field of biophysical mathematics [9]. A topology graph is created for the data where each vertex of the graph represents an unambiguous distinguishable part of the display. Figure 5 shows 5 different topological graphs. Figures 5a, b, c and d each have 9 vertices, $n = 9$. Figure 5e has $n = 11$. The edge of the graph represents a spatial or hierarchical relationship between the vertices. The degree of the vertex equals the number of edges that are incident with the given vertex. Thus, in the graph in Figure 5a, the vertex 4 is of degree two while vertex 9 is of degree one. Figures 5d and 5e are topological graphs similar to those found in visualization tools such as Narcissus [10] and Hyperbolic Trees [11].

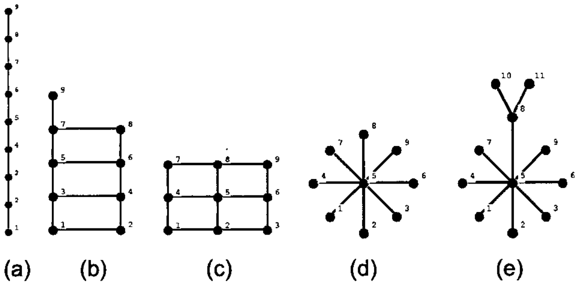


Figure 5. Topology graphs.

The vertices are partitioned into h subsets according to the degree of the vertex as seen in Table 2. A vertex is equivalent to another vertex if their degrees are equal. This is a 0th order analysis. If the equivalence relation is changed to include the first-order neighborhood, the vertices of the graphs are then partitioned as shown in Table 3. The number of vertices in the ith subset (i=1,2,...h) equals n_i and, therefore, $\Sigma n_i = n$. The subsets for the topological graphs for 0th order and 1st order equivalency are shown in Tables 2 and 3, respectively.

Table 2. Subsets for for zero-order information content.

Graph	Subset h	Degree	Vertices	n_i
a	1	1	1 and 9	2
	2	2	2, 3, 4, 5, 6, 7,and 8	7
b	1	1	9	1
	2	2	1, 2, and 8	3
	3	3	3, 4, 5, 6, 7	5
c	1	2	1, 3, 7, and 9	4
	2	3	2, 4, 6, and 8	4
	3	4	5	1
d	1	1	1, 2, 3, 4, 6, 7, 8 and 9	8
	2	8	5	1
e	1	1	1, 2, 3, 4, 6, 7, 9, 10, and 11	9
	2	3	8	1

Table 3. Subsets for first-order information content.

Graph	h	Deg	1st-order Neighborhood	Vertices	n_i
a	1	1	1-deg2	1, 9	2
	2	2	1-deg1, 1-deg2	2, 8	2
	3	2	2-deg2	3, 4, 5, 6, 7	5
b	1	1	1-deg3	9	1
	2	2	1-deg2, 1-deg3	1, 2	2
	3	2	2-deg3	8	1
	4	3	1-deg2, 2-deg3	3, 4, 6	3
	5	3	3-deg3	5	1
	6	3	1-deg1, 1-deg2, 1-deg3	7	1
c	1	2	2-deg3	1, 3, 7, 9	4
	2	3	1-deg4, 2-deg2	2, 4, 6, 8	4
	3	4	4-deg3	5	1
d	1	1	1-deg8	1, 2, 3, 4, 6, 7, 8, 9	8
	2	8	8-deg1	5	1
e	1	1	1-deg8	1, 2, 3, 4, 6, 7, 9	7
	2	1	1-deg3	10, 11	2
	3	3	2-deg1, 1-deg8	8	1
	4	8	7-deg1, 1-deg3	5	1

The probability p_i that a randomly selected vertex of the graph belongs to the *ith* subset is expressed by the relation

$$p_i = \frac{n_i}{n} \tag{13}$$

The information content based on the given equivalence relation can be calculated by Shannon's formula for entropy, which represents an average amount of information content per vertex and is given as

$$H = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{14}$$

Therefore, the topological information content for the five different layouts for zero-order and first-order equivalency is shown in Table 4.

Table 4. Topological information content for Fig. 7 graphs.

Graph	I_0 (0th order)	I_1 (1st order)
a	0.76 bits	1.44 bits
b	1.35 bits	2.42 bits
c	1.39 bits	1.39 bits
d	0.50 bits	0.50 bits
e	0.87 bits	1.49 bits

The topological information content for the graph in Figure 5d is the lowest with 0.5 bits. In general the topological information content increases as you increase the order from zero to one. The difference between 0th order and 1st order topological information content can be explained as follows. The 0th order topological information content is a concept of information content calculated from the symbols only. The 1st order (or *k*th order if need be for $k > 1$) topological information content can be regarded as a particular quantitative expression of what is created by the symbols and is a measure of their symmetry. Incidentally, the maximum topological information content for a graph with $n = 9$ nodes is $I_{max} = \log_2 9 = 3.17$, and for a graph with $n = 11$ nodes, $I_{max} = \log_2 11 = 3.46$.

4. Summary

Metrics have been developed based on Shannon's information theory to facilitate the evaluation and design of visual displays. These measures are used for quantifying the amount of information content present in the data in the form of its dimensions or its topology, and also the information content of a corresponding representational display. Again, it is emphasized that information content does not account for a semantic interpretation of the visual display. It cannot quantitatively assess any increase in information due to emergent features that were not a part of the original relation set.

5. References

[1] Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Cheshire, CT, Graphics Press.

[2] Tufte, E. R. (1990). *Envisioning Information.* Cheshire, CT, Graphics Press.

[3] Tufte, E. R. (1997). *Visual Explanations.* Cheshire, CT, Graphics Press.

[4] Bertin, J. (1981). *Graphics and Graphic Information-Processing.* New York, Walter de Gruyter.

[5] Bertin, J. (1983). *Semiology of Graphics : Diagrams, Networks, Maps.* Madison, WI, University of Wisconsin Press.

[6] Brath, R. (1997). Concept Demonstration: Metrics for Effective Information Visualization. *Proceedings IEEE Symposium on Information Visualization*, Phoenix, AZ, IEEE Computer Society.

[7] Card, S. K. and J. Mackinlay (1997). The Structure of the Information Visualization Design Space. *Proceedings IEEE Symposium on Information Visualization*, Phoenix, Arizona, IEEE Computer Society.

[8] Shannon, C. E. and W. Weaver (1964). *The Mathematical Theory of Communication.* Urbana, IL, The University of Illinois Press.

[15] Carswell, C. M. and C. D. Wickens (1988). Comparative graphics: history and applications of perceptual integrality theory and the proximity compatibility hypothesis. Urbana-Champaign, IL, Aviation Research Laboratory, University of Illinois.

[8] Mackinlay, J. (1986). "Automating the Design of Graphical Presentations of Relational Information." *ACM Transactions on Graphics* 5(2): 110-141.

[9] Rashevsky, N. (1955). "Life, information theory, and topology." *Bulletin of Mathematical Biophysics* 17: 229-235.

[10] Hendley, R. J., N. S. Drew, et al. (1995). Narcissus: Visualising Information. *Proceedings of InfoVis'95, IEEE Symposium on Information Visualization*, New York, 90-96.

[11] Lamping, J., R. Rao, and P. Pirolli (1995). A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. *CHI '95, ACM Conference on Human Factors in Computing Systems.*