Predicción del desempeño de las técnicas de visualización a partir de métricas sobre los datos*

Dana K. Urribarri
Universidad Nacional del Sur
dku@cs.uns.edu.ar

Resumen

El objetivo de una visualización es obtener una representación del conjunto de datos que ayude al usuario en la correcta interpretación de los mismos y así lograr un acertado análisis de éstos. Dado el constante crecimiento de los conjuntos de datos en diferentes y variados campos de la información, la tarea de elegir la técnica más adecuada para visualizar convenientemente los datos no es sencilla. Además, el resultado del proceso de visualización depende de todas las decisiones que se hayan tomando a lo largo de dicho proceso: un usuario inexperto es propenso a tomar decisiones equivocadas afectando negativamente la visualización obtenida y, a la larga, frustrando su experiencia con la visualización. Si bien a la hora de visualizar conjuntos de datos pequeños no hay grandes desafíos, la situación cambia al intentar visualizar grandes conjuntos de datos: una mala decisión tomada en cualquier punto del proceso de visualización y el resultado obtenido puede no ser satisfactorio. Una alternativa para solucionar este problema es guiar al usuario en la toma de decisiones a lo largo del proceso. Sin embargo, esta tarea no es sencilla: implica la existencia de herramientas que permitan predecir qué decisión es "más conveniente" tomar. Una forma de elegir la decisión más conveniente es basarse en métricas sobre los datos que describan aspectos claves de la técnica y permitan predecir el resultado final sin necesidad de aplicar la técnica sobre los datos.

1. Introducción

La principal motivación de este trabajo de tesis doctoral fue la definición de métricas asociadas a las técnicas de visualización de datos como forma de predecir qué ocurrirá con algún aspecto particular de la aplicación de dicha técnica sobre determinado conjunto de datos. De este modo se contará con un parámetro de decisión adicional a la hora de decidir con qué técnica/s visualizar cada conjunto de datos. Adicionalmente, esto permitirá definir métricas específicas que ayuden a decidir si la técnica escala visualmente para dicho conjunto de datos.

Específicamente, las contribuciones de esta tesis se agrupan en tres temas principales:

1. Diseño e implementación de un layout de árboles hiperbólico y multirresolución
2. Desarrollo de métricas que permiten establecer los límites de la escalabilidad visual en scatterplots y en árboles.
3. Integración de medidas de escalabilidad visual con el Modelo Unificado de Visualización (MUV) [19]

2. Escalabilidad Visual

La escalabilidad visual es definida por Eick y Karr [13] como la capacidad de una herramienta de visualización de mostrar efectivamente grandes conjuntos de datos, en término de la cantidad y la dimensionalidad de estos. Idealmente, la escalabilidad es cuantificable en términos de *respuestas* y *factores*:

$$\text{respuestas} = F(\text{factores}, \text{datos})$$

donde las respuestas miden el impacto del entendimiento, los descubrimientos y las decisiones inducidas por la visualización y los factores miden las características propias de la visualización.

2.1. Grandes conjuntos de datos

Las taxonomías de datos orientadas a la escalabilidad existentes en la literatura son presentadas desde un punto de vista estadístico. Tanto Unwin et al. [34] como Wegman [37] han planteado una clasificación que divide los datos según su tamaño en bytes. Sin embargo, esta clasificación tiene en cuenta solamente el tamaño de los

*Esta tesis fue presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias de la Computación, de la Universidad Nacional del Sur. Se llevó a cabo bajo la dirección de la Dra. Silvia M. Castro, Profesora Titular del Departamento de Ciencias e Ingeniería de la Computación. La tesis se defendió el 23/09/2014.

datos y por lo tanto no brinda suficiente información para elegir una técnica o estrategia de visualización acorde a los datos. Eick [12] presenta el tema particular de la escalabilidad visual de redes. Además de introducir las propiedades y la estructura de las redes a partir de la teoría de grafos, presenta algunas posibles medidas de escalabilidad visual particulares para redes: número de nodos o arcos visibles y número de elementos visibles, número de componentes conexas, entre otras.

Para el desarrollo de esta tesis se consideró que un conjunto de datos no es pequeño o grande por sí mismo, sino que su tamaño dependerá del contexto. Si el espacio disponible para graficar la visualización es reducido (por ejemplo $2cm \times 2cm$), 500 datos puede ser un conjunto *grande*. Sin embargo, la misma cantidad de datos en una visualización de $10cm \times 10cm$ puede ser un conjunto *pequeño*.

En el contexto de scatterplots Carr et al. [7] consideran que un conjunto de datos es grande cuando ocurre alguno de los siguientes escenarios: *(1)* el tiempo de "creación" de la visualización es extenso, *(2)* el tiempo de cómputo de algunas operaciones es extenso, o *(3)* la visualización tiene una gran superposición, considerando *tiempo extenso* a aquel que no sea interactivo. Extendiendo esta definición a otros contextos más generales, se puede considerar que, bajo determinadas circunstancias, un conjunto de datos es grande cuando ocurre alguno de los siguientes escenarios: *(1)* el tiempo de "creación" de la visualización es no interactivo, *(2)* el tiempo de cómputo de alguna operación es no interactivo, o *(3)* la técnica de visualización ha llegado al tope de su factor limitante.

2.2. Factores que afectan la escalabilidad visual

Los factores que la afectan la escalabilidad visual de las técnicas se dividen en dos grandes grupos: factores externos y factores internos. Los factores externos son los que no dependen de la técnica de visualización, mientras que los factores internos son los inherentes a la técnica misma.

Factores externos Eick [12] presenta diferentes factores externos que afectan la escalabilidad visual, algunos son inherentes a las personas, y otros al sistema computacional. Dentro de los factores relacionados con las personas se encuentra la *percepción humana*. Si bien el humano podría percibir unos 6,5 millones de píxeles, la *resolución del monitor* varía actualmente entre 800×600 y 2560×1600 píxeles[1] alcanzando entre 480 000 y 4 096 000 píxeles. La *interactividad* es una herramienta para incrementar la escalabilidad visual, pero se ve disminuida por la incapacidad de los usuarios para navegar espacios altamente dimensionales. Las interacciones más usuales son foco+contexto, panning y zooming, selección, agregación y brushing. También se plantean las *estructuras de datos*, los *algoritmos* empleados y la *infraestructura computacional* (cpu, red, tasa de rendering) como factores relacionados al sistema computacional que afectan la escalabilidad visual.

Factores internos Los factores internos que afectan la escalabilidad visual son aquellos factores que limitan su expresividad y están determinados por características inherentes de la técnica de visualización. En general, los factores externos anteriormente enumerados afectan a todas las técnicas de visualización, sin embargo, no todas las técnicas se ven limitadas por los mismos factores internos. Los factores internos pueden tener su origen sólo en la naturaleza de la técnica de visualización o, estar también indirectamente asociados a un factor externo. Por ejemplo, en un gráfico de barras convencional la cantidad de dimensiones representables (1 atributo por dato) está limitada por la técnica; sin embargo, no hay límite en la cantidad de barras (datos); este límite estará dado por el tamaño de la pantalla (factor externo) y el ancho de la barra (factor interno).

3. Métricas y predicción

El objetivo de una visualización es obtener una representación del conjunto de datos que ayude al usuario en la correcta interpretación de los mismos y así lograr un acertado análisis de estos. Dado el constante crecimiento de los conjuntos de datos en diferentes y variadas áreas de aplicación, la tarea de elegir la técnica más adecuada para visualizarlos convenientemente no es sencilla. Además, el resultado del proceso de visualización depende de todas las decisiones que toman a lo largo de dicho proceso: un usuario inexperto es propenso a tomar decisiones equivocadas afectando negativamente la visualización obtenida y, a la larga, frustrando su experiencia con la visualización.

Dada la gran variedad de técnicas de visualización existentes es necesario contar con medidas que ayuden a determinar qué técnica es la más adecuada para un dado conjunto de datos, así como medir la calidad o *cuán buena* es una visualización [20, 32]. Varios autores se concentran en reafirmar la necesidad y utilidad de definir métricas para caracterizar el comportamiento de las diferentes visualizaciones. En particular, algunos presentan marcos de referencia en los cuales encuadrar las métricas definidas sobre las técnicas; entre estos pueden mencionarse algunos criterios para la evaluación de técnicas de visualización ([15]), una primera sistematización de las métricas de calidad ([5]), guías para definir y comparar métricas de calidad ([31]) y un análisis de las métricas de calidad de visualizaciones de datos multi-dimensionales ([6]).

[1] Samsung 305TPlus 30" Wide

Utilizando el Modelo Unificado de Visualización [19] como marco de referencia del Pipeline de Visualización, la calidad de una visualización se podría medir a lo largo de diferentes etapas, siendo la *Vista* la etapa más directa para evaluar el resultado. Sin embargo, realizar una evaluación de la visualización a esta altura del proceso implica generar la visualización aunque esta no vaya a resultar efectiva. Nuestro objetivo es *predecir* la calidad de una visualización antes de alcanzar la *Vista*, es decir antes de aplicar la técnica al conjunto de datos.

En el proceso de visualización las medidas pueden utilizarse en diferentes etapas, tanto en la elección como en la configuración de la técnica. Llegando al final del proceso, una vez elegida la técnica de visualización y los parámetros de esta (tamaños, distancias, etc.), el usuario puede decidir, a partir del valor resultante de la evaluación de las métricas asociadas a la técnica, si la visualización sería aceptable o no para su propósito (ver figura 1).

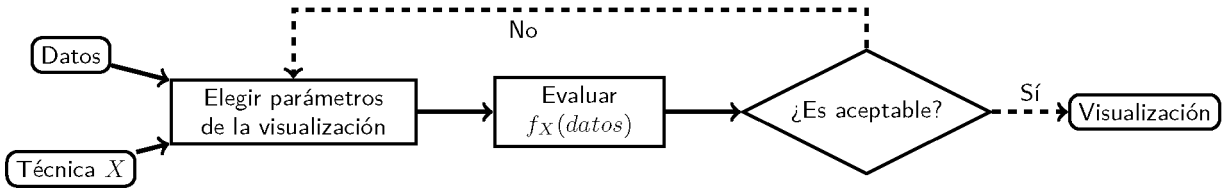

Figura 1: Proceso de refinamiento de los parámetros de la técnica. Las transiciones marcadas con una línea discontinua son aquellas que deberían contar con asistencia del usuario.

En una etapa anterior, las métricas se pueden utilizar para guiar la elección de la técnica (ver figura 2). El objetivo es evitar que el usuario intente visualizar conjuntos de datos con técnicas que, independientemente de la configuración, no darán buenos resultados pero, en cambio, sí intente con aquellas que puedan resultar en visualizaciones *potencialmente* aceptables. Esto se puede lograr, una vez elegida la técnica y antes de comenzar con el refinamiento de sus parámetros, calculando los valores límites de la métrica, es decir, llevando los parámetros que influyen en el cálculo de las medidas a los extremos que maximicen (o minimicen) su valor.

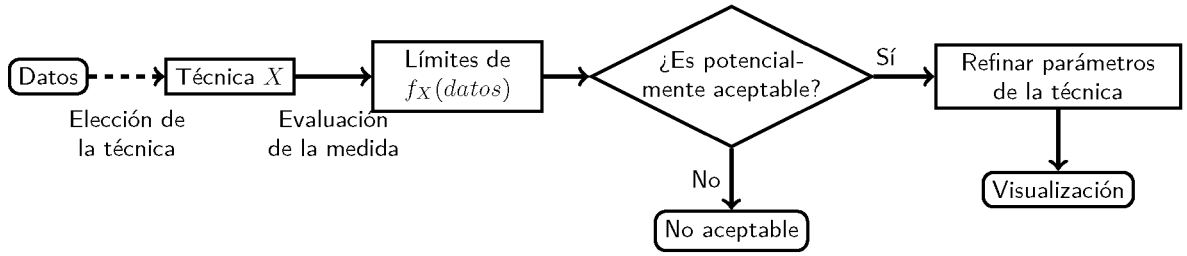

Figura 2: Proceso de elección de los parámetros de la técnica de visualización seleccionada teniendo en cuenta las métricas definidas.

Finalmente, y generalizando el caso anterior, las métricas también se pueden utilizar para determinar cuál es el conjunto de técnicas que resultarían en visualizaciones *potencialmente* aceptables para un dado conjunto de datos. De esta forma, cuando el usuario tiene que elegir una técnica para visualizar su conjunto de datos, no necesita elegir una del total de técnicas disponibles sino solamente de un subconjunto selecto (ver figura 3).

La definición de métricas asociadas a las técnicas es un elemento de gran utilidad al momento de determinar la técnica más apropiada para visualizar un determinado conjunto de datos. Sin embargo, las métricas por sí solas no son suficientes ya que estas no tienen en cuenta el tipo de dato que se está visualizando. Por lo tanto, cada vez que se propone evaluar el comportamiento de una técnica con un conjunto de datos, se asume que existió una instancia previa (por ejemplo, usando la semántica de los datos [18, 14]) en la cual se identificó esa técnica como apta para ese tipo de datos.

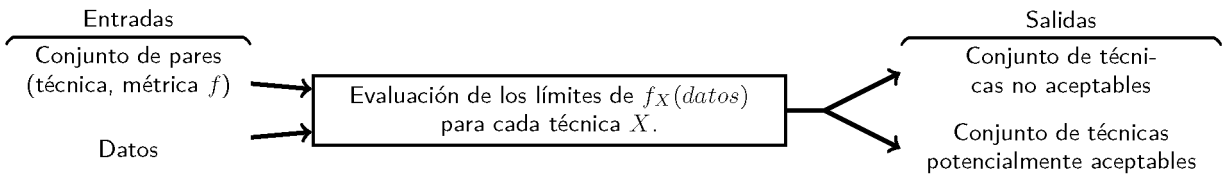

Figura 3: Dado el conjunto de datos que se desea visualizar, selección de las técnicas cuyo valor de las métricas asociadas *predicen* una visualización potencialmente aceptable.

4. Scatterplots

Un *scatterplot* (o diagrama de puntos) se define [27] como la visualización de la relación entre dos variables medidas a un mismo conjunto de individuos. Esta visualización es un diagrama que usa coordenadas cartesianas para mostrar los valores de un conjunto de datos bidimensional con un punto ($\cdot$). La posición de cada punto en el diagrama está determinada por los dos atributos del dato que se está representando. Es posible complementar el diagrama con información adicional, tal como ejes, etiquetas, leyendas o títulos y también con líneas de regresión o curvas suaves.

Los *scatterplots* tienen dos grandes limitaciones: la cantidad de dimensiones representables y la cantidad de datos que se pueden visualizar manteniendo una gráfica significativa a pesar de la superposición existente.

Dimensionalidad El *scatterplot* tradicional está limitado a 2 dimensiones. Se han propuesto diversas extensiones de la técnica para solucionar este problema y poder representar datos complejos: *scatterplots* 3D, glifos y matrices de *scatterplots*.

Superposición El mayor problema de los *scatterplots* es el alto nivel de solapamiento cuando el conjunto de datos visualizado es grande, ya que puede ocultar una parte significativa de los datos mostrados.

Estimación de densidades La inherente superposición de datos acarrea el problema de estimar la densidad de datos en cierta posición del gráfico. Los *scatterplots* representan correctamente la densidad sólo si no hay datos que se proyecten en los mismos píxeles, ya sea por la resolución limitada de la pantalla o porque hay varios datos iguales en el conjunto representado.

Como parte del trabajo previo de la tesis se realizó una recopilación de las técnicas más importantes basadas en *scatterplots* [16, 9, 10, 4, 7, 25, 3, 36, 26, 38] y de las soluciones planteadas en cada caso para obtener visualizaciones exitosas más allá de sus limitaciones. Se estudió cada una de ellas y se determinó qué problema intenta resolver cada una y mediante qué estrategia.

5. Predicción de visibilidad en scatterplots

El resultado de una visualización con *scatterplots* no depende únicamente del conjunto de datos, sino también de características particulares elegidas para la visualización: ¿cuál es el tamaño de la visualización? ¿de qué tamaño son los glifos? ¿qué forma tienen? Por otro lado, los *scatterplots* 2D son una técnica muy útil y ampliamente adoptada para la visualización de datos bidimensionales, extensible para datos multidimensionales y muy adecuada para la visualización de grandes volúmenes de datos y, dado que no existían métricas orientadas a cuantificar la escalabilidad visual de los mismos:

- Se propuso una métrica que estima los glifos siempre visibles en un *scatterplot* independientemente del orden en que estos sean dibujados.
- Se presentó una aproximación matemática de la métrica en función de la cantidad de datos a visualizar, el tamaño de la ventana y el tamaño del glifo.
- Se analizó cómo la medida definida puede asistir al usuario en la selección de los parámetros para que la visualización resulte en la *mejor vista posible* en cuanto a la visibilidad de los nodos se refiere.

El objetivo fue definir una medida que *prediga*, sin necesidad de *renderizar* la visualización, cuán *buena* será la visualización resultante. Dado que la superposición es un importante factor limitante de los *scatterplots*, se buscó una medida que exprese matemáticamente el concepto de *visibilidad*, teniendo en cuenta, además, los parámetros de la visualización $visibilidad = f(parámetros\ visualización, datos)$.

5.1. Índice de visibilidad

El *índice de visibilidad* se define como una medida específica para los *scatterplots* que, dado un conjunto de datos y las dimensiones de la ventana y del glifo, estima el porcentaje esperable de glifos que no se encuentran completamente superpuestos con otros glifos (existe al menos un píxel no superpuesto con otro glifo), es decir la cantidad esperable de glifos que serán siempre visibles independientemente del orden en que sean dibujados.

Las dimensiones de la ventana (alto y ancho) incluyen exclusivamente al área donde se grafica el *scatterplot*, es decir no incluyen menúes, bordes, botones, visualizaciones accesorias, etc. Dado que para el análisis se consideran únicamente ventanas cuadradas, con un solo valor se puede representar el alto y el ancho de la ventana y por lo tanto su tamaño. Los glifos también se consideran cuadrados y, por lo tanto, también alcanza con un único valor para representar su tamaño. En ambos casos se considera el lado del cuadrado como el tamaño del glifo o de la ventana.

Previo a modelar matemáticamente el índice de visibilidad, se desarrolló un algoritmo para calcular dicho índice dado un conjunto de n datos, el tamaño en píxeles del glifo y el tamaño de la ventana. Este algoritmo representa la visualización a través de una matriz de enteros; en la figura 4 se muestran dos ejemplos de glifos ubicados en el *scatterplot* y la matriz de enteros correspondiente a cada uno.

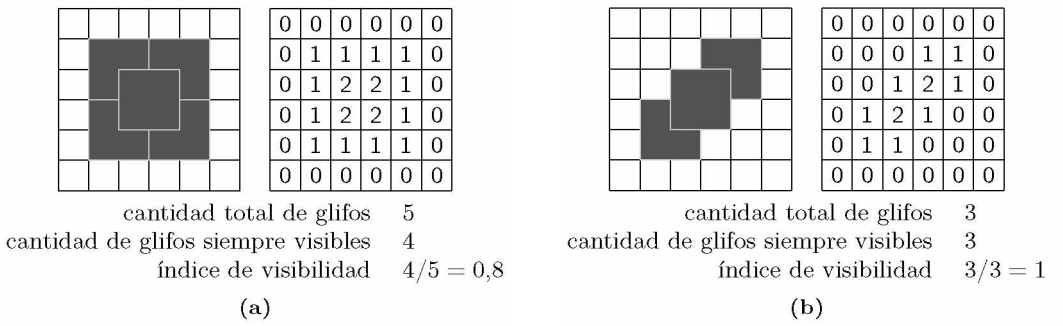

Figura 4: Ejemplos de glifos ubicados en el *scatterplot* con su matriz de enteros correspondiente. En la figura 4a se muestra un caso en el cual hay un glifo que será visible dependiendo del orden en que sea dibujado: si se dibuja primero el glifo central y luego los demás glifos, el primero quedará oculto detrás de los demás; sin embargo, si el glifo central se dibuja último es posible que sea visible, dependiendo de los colores del glifo (contorno diferente, colores diferentes, etc.). En la figura 4b se muestra un caso en el cual, independientemente del orden en que se dibujen los glifos, va a ser visible al menos un píxel de cada glifo.

5.2. Modelo matemático del índice de visibilidad

Para analizar el comportamiento de la métrica definida se experimentó con 120 conjuntos de datos generados aleatoriamente siguiendo 16 distribuciones normales diferentes para cada una de las dos dimensiones y con 23 cantidades diferentes de datos ($10; 10^{1,25}; 10^{1,5}; 10^{1,75}; 10^{2} \ldots 10^{6}; 10^{6,25}$ y $10^{6,5}$). Cada uno de estos 120 conjuntos de datos se visualizó con 375 *scatterplots* de diferentes configuraciones: 25 tamaños diferentes de ventana ($100; 300; 500 \ldots 4700$ y 4900 píxeles de lado) y 15 tamaños distintos de glifos ($2; 4; 6 \ldots 28$ y 30 píxeles de lado). Las 16 distribuciones distintas se generaron con media 1 y diferentes desvíos estándares ($0{,}05; 0{,}08; 0{,}1; 0{,}3; 0{,}5; 0{,}8; 1; 3; 5; 8; 10; 30; 50; 80; 100$ y 300).

Se utilizó la distribución normal para llevar a cabo los experimentos dado que es apropiada para la mayoría de los fenómenos naturales, cuando la muestra es muy grande y el error es lo suficientemente pequeño.

Teniendo en cuenta que el tamaño de la ventana h, el tamaño del glifo p y la cantidad de datos x son variables independientes entre sí y considerando los valores que debe cumplir la función en los límites, se aproximó el índice de visibilidad τ con la siguiente función f y se obtuvieron valores para los parámetros a, b, c y d:

$$\tau \approx f(x,h,p) = \frac{1}{1+e^{\gamma(x,h,p)}} \qquad \gamma(x,h,p) = a\ln(x) + b\ln(h) + c\ln(p) + d$$

$$a = 1{,}86056686 \quad b = -3{,}2534998 \quad c = 2{,}91520408 \quad d = -0{,}68834377$$

En todos los casos se utilizó `gnuplot` (`http://www.gnuplot.info/`) para realizar las aproximaciones, con valores iniciales en 1. `gnuplot` utiliza una implementación del algoritmo no lineal de mínimos cuadrados de Marquardt-Levenberg. Para analizar el error de la aproximación f se compararon tanto los errores absolutos como los errores máximos entre f y τ.

5.3. Cómo la métrica asiste al usuario en la configuración de un *scatterplot*

El objetivo de la métrica es guiar al usuario en la elección de los parámetros de la visualización, en particular los tamaños de la ventana y el glifo teniendo en cuenta la cantidad de datos a visualizar. Aunque las fórmulas contemplan ventanas tan grandes o glifos tan pequeños como sea necesario, en la práctica el tamaño de la ventana no debería ser mayor que el tamaño del monitor y el tamaño del glifo no puede ser menor a un píxel. En este escenario, para una dada cantidad de datos, no sería posible obtener un resultado mejor que aquél con el mayor h y el menor p posibles.

Tomando como base un monitor con una resolución máxima de 1920×1080, se presentaron dos casos de estudio que analizan cómo el *índice de visibilidad* podría ayudar al usuario en la visualización de un conjunto de datos.

En el *caso de estudio 1* el objetivo fue analizar la relación entre el tamaño del glifo y el índice de visibilidad al visualizar 1058 datos en una ventana de 400×400 píxeles. El usuario puede decidir usar glifos de un determinado tamaño o establecer una restricción en el índice de visibilidad. Si el usuario elije glifos de 16×16 píxeles, éste notaría que el índice de visibilidad es aproximadamente de 0,29. Esto implica que se espera que sólo un 29 % de los glifos tengan al menos un píxel no solapado con otro glifo. Por otro lado, si el usuario decide obtener un índice de visibilidad mínimo de 0,9, los glifos no deberían ser mayores que 5×5 píxeles.

En el *caso de estudio 2* el objetivo fue analizar la factibilidad de un scatterplot visualizando 300 000 datos. Si el usuario restringe la visualización a una ventana de 400×400 píxeles, éste notará que, incluso eligiendo

glifos de 1 píxel, el índice de visibilidad de la visualización no será mayor a 0,036. Por otro lado, si el usuario espera obtener un índice de visibilidad mínimo de 0,9, notaría que incluso con glifos de 1 píxel necesitaría una ventana mayor que lo disponible por el monitor ($h = 2155$). Más aún, si el usuario establece la ventana del máximo tamaño posible (1080×1080) y el glifo en un píxel, el índice de visibilidad seguiría siendo bajo, menor a un 50 %.

6. Visualización de árboles

La visualización de árboles se encarga de representar gráficamente estructuras jerárquicas o *árboles*, que son estructuras presentes tanto en ciencias de la computación como en otras disciplinas. Estas estructuras jerárquicas pueden contener hasta tres tipos de información de interés: *(1)* La información estructural de la jerarquía (cómo se organizan los nodos). *(2)* La información asociada a cada nodo. *(3)* La información asociada a cada enlace. Qué información es la que se desea visualizar condiciona la técnica de visualización apropiada para cada caso.

Schulz [29] establece tres propiedades que caracterizan casi cualquier técnica de visualización de árboles: *dimensionalidad* (2D, 3D o híbrido), *representación de los arcos* (explícito, implícito o híbrido), y *posicionamiento de los nodos* (radial, paralelo a ejes o libre). Dado que los árboles son un caso particular de los grafos, es esperable que la visualización de árboles tenga las mismas limitaciones que la de grafos [11]

Sobrecarga El gráfico se sobrecarga de información y se torna visualmente confuso.

Posicionamiento de los nodos La jerarquía define un orden parcial entre los elementos y la visualización debe no solo respetarlo sino también evidenciarlo.

Tensión *perceptual* La relación padre–hijo es uno de los aspectos más importantes a visualizar en una jerarquía. Además, existen relaciones entre nodos como "hermanos", "descendiente", etc. que también pueden ser relaciones importantes a distinguir en la visualización.

Se han desarrollado diferentes técnicas que compensan las limitaciones de la visualización de árboles. Los *treemaps*, *piecharts* y sus derivados son técnicas de visualización de árboles que representan de forma implícita los arcos, respetando el orden parcial existente. Además, la sobrecarga de información se evita ya que los nodos más profundos en el árbol no son representados. Sin embargo, la tensión *perceptual* puede verse perjudicada ya que los nodos que se encuentran cerca en la visualización, están cerca en la estructura, sin embargo, la inversa no siempre es válida. Los árboles hiperbólicos, mantienen la representación explícita de los arcos que respeta el orden parcial y reduce la tensión *perceptual*. Además, logran evitar la sobrecarga ya que los nodos más profundos en el árbol tampoco son representados.

Esta tesis se centró en dos técnicas de visualización de árboles, los *treemaps* como una técnica representativa de la visualización implícita de árboles y el Gyrolayout [35] como una representativa de la visualización explícita. El Gyrolayout es un layout hiperbólico interactivo de árboles desarrollado como parte de esta tesis; está basado en el layout presentado por T. Munzner [22, 23], en el Teselado baricéntrico pesado de Voronoi y los espacios gyrovectoriales de Einstein [33], que son una abstracción matemática natural para tratar con la geometría hiperbólica.

7. Predicción de la visibilidad en visualización de árboles

Al momento de visualizar grandes árboles hay que tener en cuenta si será suficiente con una técnica básica de visualización o hará falta utilizar técnicas complementarias (como, por ejemplo, *clusterización*). En lo que respecta a la gran cantidad de nodos y a la escalabilidad visual de la técnica elegida, es importante preguntarse *¿será posible representar todos los nodos, aunque sea con un único píxel?*.

Teniendo en cuenta que se busca contar con elementos que, dado un determinado conjunto de datos, permitan seleccionar una técnica de visualización adecuada, es necesario contar con medidas que predigan la calidad de la visualización. En esta tesis se desarrollaron medidas tanto para los *treemaps* como para el Gyrolayout que dado un conjunto de datos y, de ser necesario, las características propias de la visualización permiten predecir la calidad de la visualización resultante en función de la escalabilidad visual.

7.1. Predicción de la visibilidad en *treemaps*

Si bien se han desarrollado algoritmos que mejoran la calidad de la subdivisión del *treemap*, y por lo tanto mejoran la calidad de la visualización, no se han definido medidas que permitan *predecir* la calidad de esta visualización. Por lo tanto, se desarrolló una métrica particular de los *treemaps* que, en función de los parámetros de la visualización y de los datos, cuantifica la visibilidad de los nodos del árbol en el *treemap* con subdivisión *Slice and Dice* sin necesidad de renderizar la visualización.

Para desarrollar la medida se consideró que únicamente interesa visualizar la estructura del árbol; esto es, no se consideraron hojas o nodos internos con diferentes pesos o tamaños, sino que se asumió que todos las

hojas son de igual importancia (igual peso) y el peso de los nodos internos depende de la cantidad de hojas que contengan. Solamente se consideran restricciones en las dimensiones de la ventana en la cuál se realizará la visualización ($\mathbf{V} = (w\,,\,h)$), la separación entre nodos en la dirección x y en la y ($\mathbf{S} = (s_x\,,\,s_y)$) y las dimensiones mínimas requeridas para representar una hoja o un nodo interno ($\mathbf{D} = (d_x\,,\,d_y)$). Considerando las restricciones impuestas por el usuario llamaremos *porcentaje de nodos visibles* a la relación entre aquellos nodos que se representan con al menos un píxel y el total de nodos del árbol.

Para calcular el *porcentaje de nodos visibles* al visualizar el árbol $\mathcal{A}$ es necesario contar con las dimensiones $\mathbf{M}_n$ del mínimo rectángulo necesario para visualizar cada uno de los nodos del árbol. En particular, las dimensiones mínimas $\mathbf{M}$ necesarias para visualizar la raíz del árbol, son las necesarias para visualizar el árbol en su totalidad. Estas dimensiones mínimas se obtienen a través de los requerimientos $\mathbf{S}$ y $\mathbf{D}$. Sin embargo, dado que se dispone (como parte de los requerimientos) de una ventana de dimensiones $\mathbf{V}$, a un nodo n cualquiera que requiera una subventana de dimensiones mínimas $\mathbf{M}_n$ le corresponderá proporcionalmente una subventana de dimensiones $\mathbf{V}_n = (w_n\,,\,h_n)$. Luego, si $w_n \geq 1$ y $h_n \geq 1$ se tiene que el nodo n puede ser representado con al menos un píxel en una ventana de dimensiones $\mathbf{V}$. Finalmente, se calcula la proporción del total de nodos representables con al menos un píxel.

Cómo la métrica asiste al usuario en la configuración de un *Treemap* Esta métrica brinda información que le permite a un sistema semiautomático predecir cuán buena será la visualización de determinado conjunto jerárquico con un *Treemap* básico. En el caso en que la predicción no sea satisfactoria, puede sugerir cambios en los parámetros que mejoren la visualización resultante, es decir, de ser posible, sugerir disminuir la separación entre nodos, disminuir el tamaño mínimo de los mismos, y/o aumentar el tamaño de la ventana.

Si el porcentaje de nodos visibles es bajo y potencialmente mejorable, el sistema semiautomático puede sugerir cambios en los parámetros de la visualización que incrementen el valor de la medida, tales como disminuir la separación entre nodos, disminuir el tamaño mínimo deseable de los nodos o aumentar el tamaño de la ventana.

Si el porcentaje de nodos visibles no es aceptable, incluso con la mínima separación posible entre nodos ($\mathbf{S} = (1;1)$), el mínimo tamaño posible de los nodos ($\mathbf{D} = (1;1)$) y/o con la máxima ventana posible para la visualización ($\mathbf{V}$ = dimensiones del monitor), el sistema debería desestimar el *treemap* básico como una técnica potencialmente aceptable para la visualización del conjunto de datos (ver figuras 2 y 3).

7.2. Predicción de la visibilidad en el Gyrolayout

Se han definido propiedades *estéticas* de los diagramados de árboles con representación explícita de arcos [30, cap. 5] relacionadas con lograr una percepción atractiva y un diagramado visualmente eficiente. Aunque muchas veces estas propiedades estéticas sean opuestas entre sí, es deseable preservarlas para lograr una mejor lectura del árbol [8, 30, 28, 2].

Por otro lado, existen propiedades matemáticas definidas sobre grafos y otras sobre árboles en particular; además algunas de las primeras también resultan apropiadas en este último caso. Entre las más relevantes que se han tenido en cuenta en la visualización de árboles se encuentran la altura, la cantidad de nodos, hojas y nodos internos, la entropía [24, 21] o el número de Strahler [1]. Sin embargo, en la bibliografía no hay métricas de calidad orientadas a la escalabilidad visual definidas sobre visualizaciones explícitas de árboles ni tampoco sobre visualizaciones en el espacio hiperbólico (H3, Walrus, HyperbolicBrowser [17]).

En esta tesis se definió un parámetro de calidad orientado a la escalabilidad visual para evaluar las visualizaciones obtenidas utilizando el Gyrolayout y además, una métrica sobre los datos a visualizar que permite predecir el parámetro de calidad definido anteriormente.

Parámetro de calidad: visibilidad de nodos Se consideró como parámetro para medir la calidad de la visualización la cantidad de nodos visibles, es decir la *visibilidad de los nodos*. Se considerará que un árbol es *completamente visible* si todos sus nodos son visibles y *parcialmente visible* si al menos uno de sus nodos no es visible. Un nodo se considera visible si su representación en la pantalla necesita al menos un píxel. Con esto se busca establecer si existen nodos que se han acercado tanto a la superficie de la bola (el infinito del espacio hiperbólico) que no serían visibles a pesar de las posibles rotaciones de esta.

Suponiendo una ventana cuadrada de H píxeles de lado, y considerando que se aprovecha al máximo el espacio disponible, se puede estimar la cantidad de píxeles que ocupará cada nodo proyectando ortográficamente la bola de diámetro 2 que representa el espacio hiperbólico sobre tal ventana. Sea α_i el tamaño del nodo i en el espacio hiperbólico, y $A_i = \alpha_i \frac{H}{2}$ la cantidad de píxeles que ocupa el nodo en pantalla; luego, el nodo i se considera visible si $A_i \geq 1$. Un árbol será completamente visible si, para todo nodo i, $A_i \geq 1$. En caso contrario, el árbol será parcialmente visible.

Predicción de árbol completamente visible Se propuso una métrica que predice si un árbol será completamente visible o no, es decir, permite saber de antemano y *sin aplicar la técnica al conjunto de datos* si un determinado árbol se visualizará completa o parcialmente utilizando la técnica del Gyrolayout.

Para lograr este objetivo se analizaron diferentes propiedades de los árboles y se compararon con la visibilidad de árbol evaluada en una ventana de 1000×1000 píxeles. Para este análisis se estudiaron 144 árboles distintos,

entre los que se encontraban varios árboles de directorios reales y otros árboles creados al azar con características particulares (árboles completos de diferentes grados y alturas, árboles con poca cantidad de nodos y mucha altura, y árboles con muchos nodos y poca altura). En particular, la relación *cantidad de nodos–altura* fue encontrada como una de las más prometedoras a la hora de identificar árboles completamente visibles: al graficar el logaritmo de la cantidad de nodos contra el logaritmo de la altura (ver figura 5), se puede apreciar que aquellos árboles que se ubican debajo de la línea ℓ son, en general, árboles completamente visibles. En los experimentos se logró durante la predicción de árboles completa y parcialmente visibles un 88 % de aciertos contra un 12 % de predicciones erróneas.

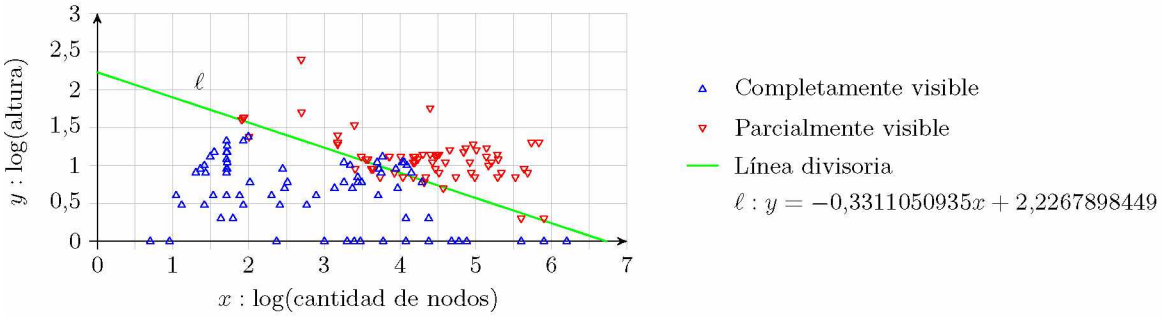

Figura 5: Predicción de la visibilidad de nodos en el Gyrolayout.

Asistencia al usuario A partir de la métrica propuesta, un sistema semiautomático puede tener en cuenta que si el árbol se ubica por debajo de la línea divisoria, será completamente visible y si el árbol se ubica por encima de la línea divisoria, será parcialmente visible. Sin embargo, si el árbol se ubica en la cercanía de la línea divisoria, no es posible saber si será completa o parcialmente visible.

Igualmente, y a pesar de la incertidumbre existente, es posible que para una gran cantidad de árboles, la métrica ayude al sistema (y por lo tanto al usuario) a predecir correctamente el desempeño del Gyrolayout a la hora de visualizar determinado árbol de datos.

8. Conclusiones

Con el objetivo de contribuir en la toma de decisiones a lo largo del pipeline de visualización, en esta tesis se ha propuesto la aplicación de métricas específicas de escalabilidad visual para técnicas de visualización representativas con el propósito de *predecir* cómo se desempeñarán estas en función de sus factores limitantes para un determinado conjunto de datos.

A partir de la división de los factores limitantes de las técnicas en factores externos e internos, se propuso la definición de métricas orientadas a la escalabilidad visual que cuantifiquen los factores limitantes internos de estas. Dado que, en general, un gran factor limitante de las técnicas de visualización es la oclusión o la superposición debido al reducido espacio disponible para realizar las visualizaciones, se definieron medidas que permiten predecir la *visibilidad* de los datos en la visualización. En particular, se propusieron métricas para predecir la visibilidad en los *scatterplots*, en los *Treemaps* y en el Gyrolayout y se propuso la integración de estas métricas al proceso de visualización. Adicionalmente, se desarrolló un *layout* particular para la visualización de árboles. Específicamente las contribuciones se pueden agrupar en:

Gyrolayout Se desarrolló una implementación novedosa de un *layout* hiperbólico 3D con capacidades multirresolución. Si bien inicialmente se representan todos los nodos, cuando la cantidad de estos es grande y ofuscan la vista es posible colapsar subárboles para reducir la cantidad de datos visibles tratando de conservar la información. Como parte del desarrollo del Gyrolayout se propuso una extensión del Teselado baricéntrico pesado de Voronoi para subdividir la superficie de una esfera.

Escalabilidad visual en *scatterplots* y árboles Se desarrollaron métricas específicas de escalabilidad visual tanto para *scatterplots* como para árboles.

En el caso de los *scatterplots* se definió una fórmula que modela la visibilidad de datos distribuidos según una distribución normal en un *scatterplot* en función de la cantidad de datos, el tamaño de la ventana y el tamaño del glifo; se considera tanto la ventana como el glifo cuadrados.

En lo referido a los árboles se consideraron dos casos, los *treemaps* y el Gyrolayout. En el caso de los *treemaps* se presentó un algoritmo que calcula el porcentaje de nodos visibles según las propiedades definidas para la visualización (dimensiones mínimas de los nodos, separación entre nodos y dimensiones de la ventana). Para el Gyrolayout se propone una propiedad del árbol como métrica para predecir la visibilidad de sus nodos

en la visualización, permitiendo distinguir entre aquellos árboles que serán totalmente visibles de aquellos parcialmente visibles.

Integración de medidas de escalabilidad visual con el Modelo Unificado de Visualización Se integró la propuesta de utilización de métricas al MUV, permitiendo asistir al usuario en la etapa de elección de la técnica más apropiada para un dado conjunto de datos.

La definición de medidas que predigan algún factor importante de las técnicas de visualización en función del conjunto de datos y de las características particulares de su instanciación, permite predecir aspectos del resultado final de la técnica y por lo tanto distinguir entre aquellas potencialmente efectivas y aquellas que no lo son. Además, estas le permiten al usuario encontrar una configuración aceptable de la técnica sin necesidad de concretar la visualización, incluyendo las capacidades multirresolución de las técnicas.

En lo que al desarrollo de nuevas técnicas de visualización se refiere, esta tesis plantea la necesidad de que estas estén acompañadas de un análisis de sus factores limitantes internos y de métricas propias que permitan predecir su desempeño y estimar los parámetros más apropiados para visualizar un dado conjunto de datos.

9. Trabajo Futuro

En esta tesis se planteó la necesidad de que cada técnica de visualización tuviera asociado un conjunto de métricas que indicaran cómo se desempeña la técnica con un determinado conjunto de datos en relación a sus factores limitantes. En base a esto es posible plantear dos tipos de trabajos futuros, aquellos que completan el trabajo aquí presentado y aquéllos derivados que lo extienden:

Trabajos futuros complementarios Las tres métricas que se presentaron a lo largo de los diferentes capítulos son métricas asociadas a casos particulares de cada técnica, por lo tanto se propone extender la métrica definida para los *scatterplots* a otras distribuciones de datos además de la distribución normal, a ventanas no cuadradas y a glifos de diferentes formas; extender la métrica definida para los *treemaps* a casos en los que haya información asociada a cada nodo y la estructura del árbol no sea lo único importante a visualizar, es decir, que el tamaño de cada nodo sea proporcional a algún dato asociado; y formalizar una métrica para la visibilidad en el Gyrolayout que complete la propuesta presentada, la extienda a diferentes tamaños de ventanas y permita determinar el porcentaje de visibilidad de los nodos.

Trabajos futuros derivados Dado que el trabajo realizado en esta tesis está enfocado sobre tres técnicas representativas de visualización, para cada una de las cuales se definió una única métrica, se propone definir nuevas métricas que midan el desempeño de las técnicas expuestas en esta tesis en función de otros factores limitantes, y analizar los factores limitantes de otras técnicas de visualización ya existentes para luego definir métricas que midan el desempeño de estas técnicas en función de sus factores limitantes.

Referencias

[1] D. Auber, M. Delest, J. P. Domenger, P. Duchon y J. M. Fédou. "New Strahler numbers for rooted plane trees". En: *Third Colloquium on Mathematics and Computer Science*. 2004, págs. 203-215.

[2] Giuseppe Di Battista, Peter Eades, Roberto Tamassia y Ioannis G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999. ISBN: 0133016153.

[3] Barry G. Becker. *Volume Rendering for Relational Data*. Inf. téc. Silicon Graphics Inc., 1997.

[4] Richard A. Becker y William S. Cleveland. "Brushing Scatterplots". En: *Technometrics* 29.2 (1987), págs. 127-142.

[5] Enrico Bertini y Giuseppe Santucci. "Visual quality metrics". En: *Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization*. BELIV '06. Venice, Italy: ACM, 2006, págs. 1-5. ISBN: 1-59593-562-2.

[6] Enrico Bertini, Andrada Tatu y Daniel Keim. "Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization". En: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (dic. de 2011), págs. 2203-2212. ISSN: 1077-2626.

[7] D. B. Carr, R. J. Littlefield, W. L. Nicholson y J. S. Littlefield. "Scatterplot matrix techniques for large N". En: *Journal of the American Statistical Association* 82.398 (1987), págs. 424-436.

[8] Timothy M. Chan, Michael T. Goodrich, S. Rao Kosaraju y Roberto Tamassia. "Optimizing Area and Aspect Ratio in Straight-Line Orthogonal Tree Drawings". En: *Graph Drawing*. Ed. por Stephen C. North. Vol. 1190. Lecture Notes in Computer Science. Springer, 1996, págs. 63-75. ISBN: 3-540-62495-3.

[9] Herman Chernoff. "The Use of Faces to Represent Points in K-Dimensional Space Graphically". En: *Journal of the American Statistical Association* 68.342 (1973), págs. 361-368.

[10] William S. Cleveland y Robert McGill. "The Many Faces of a Scatterplot". En: *Journal of the American Statistical Association* 79.388 (1984), págs. 807-822.

[11] Stephen G. Eick. "Aspects of Network Visualization". En: *IEEE Computer Graphics and Applications* 16.2 (1996), págs. 69-72. ISSN: 0272-1716.

[12] Stephen G. Eick. "Scalable Network Visualization". En: ed. por Charles D. Hansen y Christopher R. Johnson. Elsevier Academic Press, 2005. Cap. 42, págs. 819-829. ISBN: 0-12-387582-X.

[13] Stephen G. Eick y Alan F. Karr. *Visual Scalability.* Inf. téc. National Institute of Statistical Sciencies, 2000.

[14] Sebastian Escarza, Martin L. Larrea, Dana K. Urribarri, Silvia M. Castro y Sergio R. Martig. "Integrating Semantics into the Visualization Process". En: *Scientific Visualization: Interactions, Features, Metaphors.* Ed. por Hans Hagen. Vol. 2. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2011, págs. 92-102. ISBN: 978-3-939897-26-2.

[15] Carla M. D. S. Freitas, Paulo R. G. Luzzardi, Ricardo A. Cava, Marco Winckler, Marcelo S. Pimenta y Luciana P. Nedel. "On evaluating information visualization techniques". En: *Proceedings of the Working Conference on Advanced Visual Interfaces.* AVI '02. Trento, Italy: ACM, 2002, págs. 373-374. ISBN: 1-58113-537-8.

[16] Daniel A. Keim, Ming C. Hao, Umeshwar Dayal, Halldor Janetzko y Peter Bak. "Generalized Scatter Plots". En: *Information Visualization* 9.4 (2010), págs. 301-311.

[17] John Lamping, Ramana Rao y Peter Pirolli. "A focus+context technique based on hyperbolic geometry for visualizing large hierarchies". En: *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems.* Denver, Colorado, United States: ACM Press/Addison-Wesley Publishing Co., 1995, págs. 401-408.

[18] Martín L. Larrea. "Visualización Basada en Semántica". Tesis doct. Universidad Nacional del Sur, 2010.

[19] Sergio Martig, Silvia Castro, Pablo Fillottrani y Elsa Estevez. "Un Modelo Unificado de Visualización". En: *IX Congreso Argentino de Ciencias de la Computación.* La Plata, Argentina, 2003.

[20] Nancy Miller, Beth Hetzler, Grant Nakamura y Paul Whitney. "The need for metrics in visual information analysis". En: *Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation.* NPIV '97. Las Vegas, Nevada, United States: ACM, 1997, págs. 24-28. ISBN: 1-58113-051-1.

[21] Abbe Mowshowitz y Matthias Dehmer. "Entropy and the Complexity of graphs revisited". En: *Entropy* 3.14 (2012), págs. 559-570.

[22] Tamara Munzner. "H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space". En: *IEEE Symposium on Information Visualization.* 1997, págs. 2-10.

[23] Tamara Munzner. "Interactive Visualization of Large Graphs and Networks". Tesis doct. Stanford University, 2000.

[24] Roberto Navigli y Mirella Lapata. "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation". En: Hyderabad, India, 2007, págs. 1683-1688.

[25] R.M. Pickett y G.G. Grinstein. "Iconographic Displays For Visualizing Multidimensional Data". En: *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics, 1988.* Vol. 1. 1988, págs. 514-519.

[26] Harald Piringer, Robert Kosara y Helwig Hauser. "Interactive Focus+Context Visualization with Linked 2D/3D Scatterplots". En: *Proceedings of the Second International Conference on Coordinated & Multiple Views in Exploratory Visualization.* CMV '04. Washington, DC, USA: IEEE Computer Society, 2004, págs. 49-60. ISBN: 0-7695-2179-7.

[27] John Renze y Eric W. Weisstein. *"Scatter Diagram." From MathWorld–A Wolfram Web Resource.*

[28] Adrian Rusu y Confesor Santiago. "Grid drawings of binary trees: An experimental study". En: *Journal of Graph Algorithms and Applications* 12.2 (2008), págs. 131-195.

[29] Hans-Jörg Schulz. "Treevis.net: A Tree Visualization Reference". En: *Computer Graphics and Applications, IEEE* 31.6 (2011), págs. 11-15.

[30] Roberto Tamassia, ed. *Handbook of Graph Drawing and Visualization.* CRC Press, 2013.

[31] Andrada Tatu, Peter Bak, Enrico Bertini, Daniel Keim y Joern Schneidewind. "Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data". En: *Proceedings of the International Conference on Advanced Visual Interfaces.* AVI '10. Roma, Italy: ACM, 2010, págs. 49-56. ISBN: 978-1-4503-0076-6.

[32] Edward R. Tufte. *The Visual Display of Quantitative Information.* 2da edición (1era edición, 1983). Graphics Press, 2001.

[33] Abraham A. Ungar. *Analytic Hyperbolic Geometry. Mathematical Foundations and Applications.* World Scientific Publishing Co. Pte. Ltd, 2005.

[34] Antony Unwin, Martin Theus y Heike Hofmann. *Graphics of Large Datasets: Visualizing a Million (Statistics and Computing).* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN: 0387329064.

[35] Dana K. Urribarri, Silvia M. Castro y Sergio R. Martig. "Gyrolayout: A Hyperbolic Level-of-Detail Tree Layout". En: *Journal of Universal Computer Science* 19.1 (2013), págs. 132-156.

[36] Martin Wattenberg y Danyel Fisher. "A model of multi-scale perceptual organization in information graphics". En: *Proceedings of the Ninth annual IEEE conference on Information visualization.* INFOVIS'03. Seattle, Washington: IEEE Computer Society, 2003, págs. 23-30. ISBN: 0-7803-8154-8.

[37] E. Wegman. "Huge Data Sets and the Frontiers of Computational Feasibility". En: *Journal of Computational and Graphical Statistics* 4 (1995), págs. 281-295.

[38] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer New York, 2009. ISBN: 978-0-387-98140-6.