Sobre coleções e aspectos de centralidade em dados multidimensionais

Douglas Cedrim Oliveira

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Douglas Cedrim Oliveira

Sobre coleções e aspectos de centralidade em dados multidimensionais

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Antonio Castelo Filho

USP – São Carlos
Outubro de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

O389s

Oliveira, Douglas Cedrim
Sobre coleções e aspectos de centralidade em dados multidimensionais / Douglas Cedrim Oliveira; orientador Antonio Castelo Filho. - São Carlos - SP, 2016.
137 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. Visualização de Informação; Projeção Multidimensional; Redução de Dimensionalidade; Nuvens de Palavras; Visualização de Texto; Medidas de Qualidade; Funções de Profundidade de Dados; Estatística Não-paramétrica. I. Filho, Antonio Castelo, orient. II. Título.

Douglas Cedrim Oliveira

On collections and centrality aspects of multidimensional data

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Antonio Castelo Filho

USP – São Carlos
October 2016

AGRADECIMENTOS

Agradeço ao Prof. Antônio Castelo Filho pela aceitação de orientação no doutorado. Além disso, aos professores Afonso Paiva Neto, Luis Gustavo Nonato e Leandro Franco de Souza. Suas contribuições ao longo desses anos, tanto acadêmicas/científicas quanto pessoais foram muito importantes.

Agradeço aos professores e pesquisadores: Fabiano Petronetto, João Paulo Gois, Mario Liziér e Wallace Casaca, pela sua disponibilidade em constituírem a banca de defesa dessa tese de doutorado.

Agradeço também ao Instituto de Ciências Matemáticas e da Computação da Universidade de São Paulo (ICMC - USP) por toda a infraestrutura disponibilizada, não somente através do Laboratório de Matemática Aplicada e Computação Científica (LMACC) mas também por toda a estrutura de biblioteca e de convivência harmoniosa com os outros estudantes e funcionários.

A cada um daqueles que integrou os grupos de Visualização e Processamento Geométrico (*VGPG*) e o LMACC, ao longo do período 2011-2016, ainda que não esteja mais entre nós... Independente de nacionalidades, seja do Peru, Colômbia, Chile, Paraguai, Alemanha, Paquistão ou Brasil, a convivência cotidiana com todos marca um grande aprendizado pessoal e acadêmico para mim, culminando em boas relações profissionais e de amizade. Em particular, agradeço aos amigos Josuel Kruppa, Erick Gomez-Nieto, Vinícius Borges e Adriano Takata, tanto pela ajuda quanto pelo apoio diário nos momentos finais desse período - um dos mais difíceis.

Agradeço também à Fapesp pelo apoio financeiro tanto no projeto de doutorado, processo número 2011/12263-0 quanto no estágio BEPE processo número 2014/11296-0 que possibilitaram conduzir essa pesquisa.

O estágio no grupo de visualização da Technische Universität Wien (TU Wien) sob orientação do prof. M. Eduard Gröller foi importante para o desenvolvimento desse trabalho. Além de um inestimável aprendizado pessoal e científico, possibilitou também o início da colaboração mais efetiva com aquele grupo de pesquisa.

Por fim, porém não menos importante, um agradecimento especial à minha família de sangue, e à construída ao longo do caminho (amigos mais próximos), que tanto vêm torcendo pela minha evolução acadêmica ao longo dos últimos anos. Em particular, à Camila Alice por seu constante apoio nesses anos e à minha mãe Yêda, pelo seu imenso carinho e amor.

“Ligar pontos pode ser tão difícil quanto divertido… ”

(Anônimo)

RESUMO

CEDRIM, D.. **Sobre coleções e aspectos de centralidade em dados multidimensionais**. 2016. 137 f. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A análise de dados multidimensionais tem sido por muitos anos tópico de contínua investigação e uma das razões se deve ao fato desse tipo de dados ser encontrado em diversas áreas da ciência. Uma tarefa comum ao se analisar esse tipo de dados é a investigação de padrões pela interação em projeções multidimensionais dos dados para o espaço visual. O entendimento da relação entre as características do conjunto de dados (*dataset*) e a técnica utilizada para se obter uma representação visual desse *dataset* é de fundamental importância uma vez que esse entendimento pode fornecer uma melhor intuição a respeito do que se esperar da projeção. Por isso motivado, no presente trabalho investiga-se alguns aspectos de centralidade dos dados em dois cenários distintos: coleções de documentos com grafos de coautoria; dados multidimensionais mais gerais.

No primeiro cenário, o dado multidimensional que representa os documentos possui informações mais específicas, o que possibilita a combinação de diferentes aspectos para analisá-los de forma sumarizada, bem como a noção de centralidade e relevância dentro da coleção. Isso é levado em consideração para propor uma metáfora visual combinada que possibilite a exploração de toda a coleção, bem como de documentos individuais.

No segundo cenário, de dados multidimensionais gerais, assume-se que tais informações não estão disponíveis. Ainda assim, utilizando um conceito de estatística não-paramétrica, denominado funções de profundidade de dados (*data-depth functions*), é feita a avaliação da ação de técnicas de projeção multidimensionais sobre os dados, possibilitando entender como suas medidas de profundidade (centralidade) foram alteradas ao longo do processo, definindo uma também medida de qualidade para projeções.

Palavras-chave: Visualização de Informação; Projeção Multidimensional; Redução de Dimensionalidade; Nuvens de Palavras; Visualização de Texto; Medidas de Qualidade; Funções de Profundidade de Dados; Estatística Não-paramétrica.

ABSTRACT

CEDRIM, D.. **Sobre coleções e aspectos de centralidade em dados multidimensionais**. 2016. 137 f. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Analysis of multidimensional data has been for many years a topic of continuous research and one of the reasons is such kind of data can be found on several different areas of science. A common task analyzing such data is to investigate patterns by interacting with spatializations of the data onto the visual space. Understanding the relation between underlying dataset characteristics and the technique used to provide a visual representation of such dataset is of fundamental importance since it can provide a better intuition on what to expect from the spatialization. Motivated by this, in this work we investigate some aspects of centrality on the data in two different scenarios: document collection with co-authorship graphs; general multidimensional data.

In the first scenario, the multidimensional data which encodes the documents is much more information specific, meaning it makes possible to combine different aspects such as a summarized analysis, as well as the centrality and relevance notions among the documents in the collection. In order to propose a combined visual metaphor, this is taken into account make possible the visual exploration of the whole document collection as well as individual document analysis.

In the second case, of general multidimensional data, there is an assumption that such additional information is not available. Nevertheless, using the concept of data-depth functions from non-parametric statistics it is analyzed the action of multidimensional projection techniques on the data, during the projection process, in order to make possible to understand how depth measures computed in the data have been modified along the process, which also defines a quality measure for multidimensional projections.

Key-words: Information Visualization; Multidimensional Projection; Dimensionality Reduction; Word Clouds; Text Visualization; Quality Measures; Data-Depth Functions; Non-parametric Statistics.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

L_1D ..... Função de profundidade da mediana geométrica
ARF ..... *Attractive and Repulsive Forces*
CHD ..... Função de profundidade baseada em fecho convexo (*Convex-Hull peeling Depth*)
FA2 ..... ForceAtlas2
FR ..... Fruchterman-Reingold
GMHD ... *Generalized Mahalanobis Depth*
i.i.d. ..... identicamente e independentemente distribuídos
IDF ..... *Inverse Document Frequency*
k-NN ..... *k−nearest neighbors*
KDD ..... *Knowledge Discovery in Databases*
kmGMHD *kernel Generalized Mahalanobis Depth*
KPCA .... *Kernel Principal Component Analysis*
LAMP ... *Local Affine Multidimensional Projection*
LDA ..... *Latent Dirichlet Allocation*
LSP ..... *Least Square Projection*
MHD ..... Função de profundidade de *Mahalanobis*
MIST .... *Multiscale Information and Summaries of Texts*
MP ..... Multidimensional Projection
PCA ..... *Principal Component Analysis*
SPLOM .. Matriz de gráficos de dispersão (*Scatter Plot Matrix*)
TF ..... *Term Frequency*
TF-IDF ... *Term Frequency-Inverse Document Frequency*
YH ..... Yifan Hu

LISTA DE SÍMBOLOS

X — Conjunto ordinário de pontos

$\mathbf{x}$ — Ponto multidimensional / vetor

$\mathbb{R}^m$ — Espaço euclidiano m-dimensional

$\mathbf{M}$ — Matriz ordinária

F_n — Distribuição empírica de probabilidade de um conjunto com n amostras

F — Distribuição de probabilidade

μ_{F_n} — Média associada a uma distribuição empírica de probabilidade

$\mathbf{C}_{F_n}$ — Matriz de covariância amostral

D^m — Campo esalar de profundidade de dados definida em pontos m-dimensionais

$\mathscr{H}$ — Espaço de Hilbert

κ — Função *kernel*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A importância dada a análise de dados tem crescido enormemente nos últimos anos. Diferentes razões fazem com que a evolução dessas análises seja considerada uma tarefa desafiadora. Primeiramente, as fontes de dados são atualmente ubíquas e estão disponíveis para uma grande audiência. Essa facilidade pode ser justificada, por exemplo, pela redução do custo monetário atribuído a sensores de alta definição como câmeras e *scanners* 3D. O volume de dados gerados por usuários em uma pequena escala de tempo (por ex., horas) pode facilmente atingir uma escala de *gigabytes* (ou *terabytes*). Segundo Keim e Ward (2002), o aumento no volume de dados disponíveis em forma digital por ano é da ordem de um milhão de *terabytes*, assim, extrair informações que sejam relevantes de tamanho volume de dados se torna uma tarefa de grande importância. E, por último, a complexidade dos dados em si também é um importante aspecto, pela possível heterogeneidade dos dados, que assumem as mais variadas formas como bancos de dados genéticos (SEO; SHNEIDERMAN, 2005), simulações computacionais (BERGER *et al.*, 2011), coleções de imagens (JOIA *et al.*, 2011), e dados textuais (Van Der Maaten; HINTON, 2012; GOMEZ-NIETO *et al.*, 2014).

Na área de *Aprendizado de Máquina* é comum encontrar várias abordagens que possibilitam análise dos dados de formas mais automáticas, sem necessariamente utilizar a intervenção do usuário no processo. Nesse contexto, relações devem ser aprendidas a partir dos dados de entrada com o objetivo de otimizar alguma medida de performance. Para isso, é necessário que haja um conhecimento prévio mínimo sobre os dados, por exemplo a informação de classes distintas as quais alguns grupos de dados de entrada pertencem. Assim, hipóteses são geradas a partir dessa informação, de forma que consigam discriminar regiões importantes no espaço de entrada, por exemplo separando grupos de dados de diferentes classes (KULIS, 2013). Vale observar que hipóteses geradas sobre os dados estão diretamente relacionadas à medida de performance, de forma que não necessariamente *insights* sobre os dados são obtidos. Embora algoritmos que realizam essas operações de maneira automática sejam bastante úteis, eles são altamente dependentes da definição de medidas (ou métricas) de similaridade adequadas entre todos os

pares dos dados de entrada, o que motiva a intervenção do usuário no processo (KULIS, 2013).

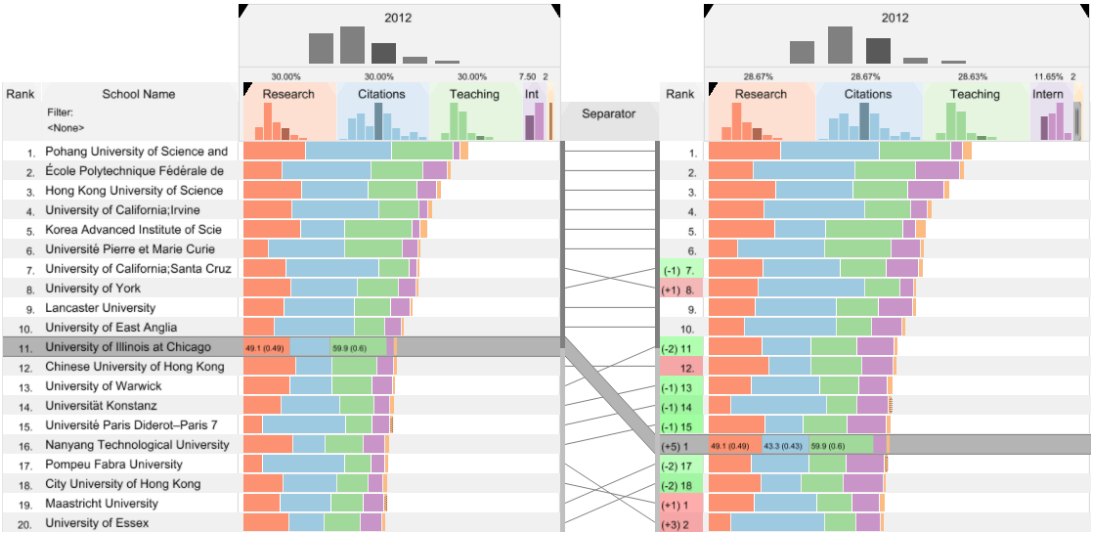
Diretamente relacionada está a área de Descoberta de conhecimento em bancos de dados - *Knowledge Discovery in Databases* (KDD). Segundo Fayyad *et al.* (1996), o processo de *KDD* pode ser definido pelo interesse em que padrões sejam extraídos a partir dos dados brutos de forma algorítmica. Note que isso implica em um um detalhe fundamental se comparado à aprendizado de máquina: extrair conhecimento dos dados eventualmente *sem* informação prévia (e.g. determinar classes a partir do dado bruto). De forma geral, as etapas envolvidas no processo são: Seleção, Pré-processamento, Transformação, Mineração, Interpretação / Avaliação (FAYYAD *et al.*, 1996).

A área de Visualização de Informação relaciona-se com esse processo de *KDD* de formas variadas: a) através da representação visual dos padrões extraídos dos dados, tornando-os de mais fácil interpretação para analistas; b) através da Análise Exploratória de Dados, onde o usuário tem papel central em guiar tarefas de Mineração dos dados, interagindo e modificando os dados através de metáforas visuais como gráficos de dispersão (*scatter plots*), tendo como o principal objetivo buscar padrões, estruturas e tendências nos dados (OLIVEIRA; LEVKOWITZ, 2003). Um exemplo é ilustrado na Figura 1, onde a abordagem possibilita a exploração visual de *rankings*, baseados em algum atributo específico, usando metáfora visual híbrida de tabelas e gráficos de barra. O processo de exploração visual para construção do conhecimento é interativo em geral. Por exemplo, o usuário poder refazer um ranking através da criação de novos atributos derivados dos atributos originais, através de combinação com pesos de cada um deles, de modo a explorar como o ranking original é modificado. No exemplo ilustrado, o peso relativo ao atributo de estágio (*internship*) é aumentado, mostrando que caso em um novo ranking fosse dado uma maior relevância a esse atributo a Universidade de Illinois perderia cinco posições. Mostra também que as seis primeiras universidades do ranking original não teriam suas posições modificadas.

1.1 Dados multidimensionais

Ao falar de atributos no exemplo da Figura 1 estamos implicitamente falando de dimensões que caracterizam o dado. De forma simplificada, os dados que discutiremos ao longo deste trabalho são ditos *multidimensionais*, isto é, um conjunto X de *dados multidimensionais* é definido como $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ tal que $\mathbf{x}_i \in \mathbb{R}^m, i = 1, \ldots, n$ e $m \geq 4$. Para $m = 1, 2, 3$ utilizaremos a notação de dados uni, bi e tridimensionais respectivamente. Oliveira e Levkowitz (2003) discutem que a distinção entre dados multidimensionais de baixa e alta dimensão não é tão precisa, entretanto caracteriza baixa dimensionalidade ($1 \leq m \leq 4$), média dimensionalidade ($5 \leq m \leq 9$) e alta dimensionalidade ($m \geq 10$). Adicionalmente, Beyer *et al.* (1999) estimam que, dependendo de suas propriedades, dados em espaços com dimensão $10 \leq m \leq 20$ já podem tornar algoritmos que dependam de vizinhança instáveis.

Figura 1 – Exploração visual de *rankings*. O ranking original é exibido à esquerda e após a modificação dos pesos é comparado com o original à direita, enfatizando as mudanças de posição cada elemento.



Fonte: Gratzl *et al.* (2013, Página 6).

Em cada uma dessas dimensões os dados podem variar de acordo quanto à sua natureza: contínua (número real), discreta (número inteiro); e quanto ao seu tipo: quantitativos (numéricos) e qualitativos (categóricos). Observe que um paralelo pode ser feito com o contexto de banco de dados, onde o conjunto de dados multidimensionais X é uma tabela e cada ponto multidimensional de dimensão m do conjunto é uma $m-tupla$.

A menos que seja explicitamente mencionado, ao longo desse trabalho trabalharemos com dados multidimensionais numéricos discretos, sejam inteiros ou reais representados em ponto flutuante.

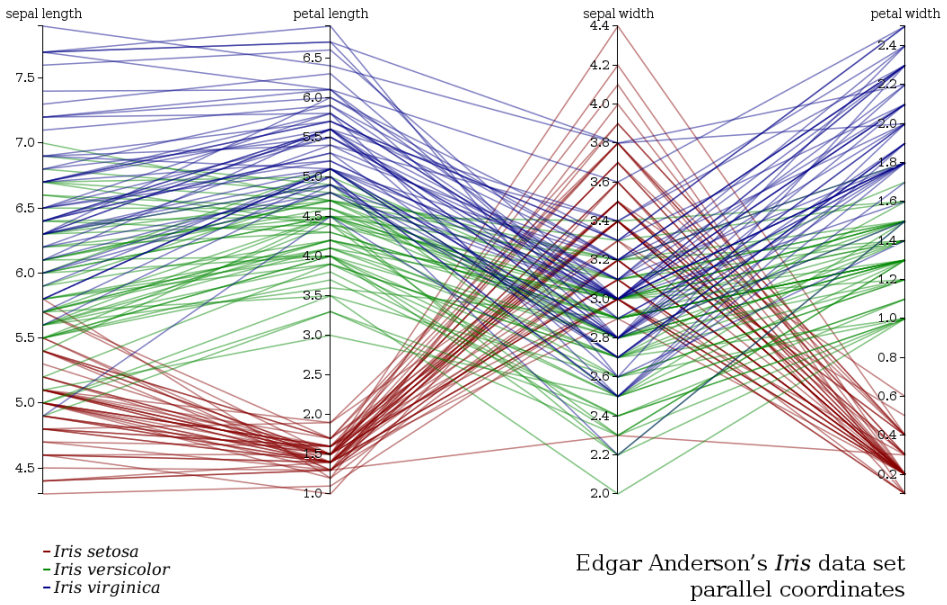
Internamente à comunidade de visualização, são muitos os esforços destinados a prover ferramentas e técnicas que permitam uma análise de dados multidimensionais. Keim (1997) classifica as abordagens como: geométricas, baseadas em ícones, orientadas a pixel, hierárquicas, baseadas em grafo e híbridas, ou como uma combinação das anteriores. Além disso, classifica também as técnicas quanto à sua forma de distorção e de interação com usuário.

Cada uma dessas categorias possui seus objetivos quanto ao comportamento que deseja-se analisar. Dentro da categoria geométrica, por ex., cada dado multidimensional é associado à uma entidade geométrica: ponto (*scatter plots*), linhas (coordenadas paralelas). Os objetivo da análise, as perguntas que querem ser investigadas pelo usuário, acabam guiando a escolha de uma metáfora em particular.

Coordenadas paralelas, por ex., utilizam que cada dimensão defina um eixo (usualmente vertical), que em conjunto estarão organizados em paralelo, onde uma linha cortando esses eixos corresponde a um ponto multidimensional, cujas coordenadas definem as intersecções da linha com os eixos. É uma metáfora que possibilita visualizar tanto as coordenadas dos dados no

espaço original quanto como as dimensões relacionam-se entre si, como ilustrado na Figura 2. Uma limitação dessa abordagem consiste em que não há uma ordenação única para o arranjo dos eixos, de forma que analisar a relação entre as dimensões e os padrões extraídos disso fica dependente de uma ordenação adequada de cada dimensão. Além disso, perde-se a noção de distância entre conjuntos de pontos.

Figura 2 – Coordenadas paralelas de dados de dimensão quatro, onde cada linha (ponto em $\mathbb{R}^4$) é representa as medidas de largura e comprimento da pétala e da sépala de flores do gênero Íris. As diferentes cores refletem as três espécies visualizadas.

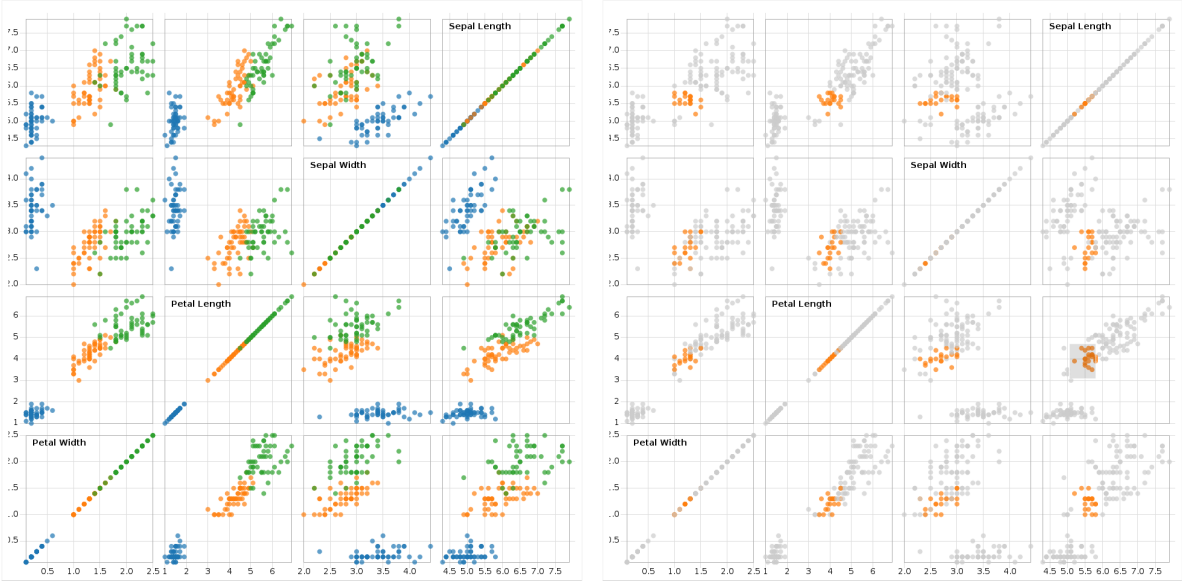


Fonte: Adaptada de Bostock (2012).

Por outro lado, o uso de gráficos de dispersão (*scatter plots*), que utiliza eixos cartesianos ortogonais, é comumente utilizado para visualizar a noção de localização, agrupamentos (*clusters*) e distância entre os seus elementos. Entretanto, seu uso fica limitado em dados de dimensão $m \leq 3$. Uma forma de contornar essa limitação é através da utilização de matrizes de gráficos de dispersão (SPLOM), onde são gerados gráficos de dispersão bidimensionais para cada par de dimensões com os pontos projetados ortogonalmente naquela direção, como ilustrado na Figura 3a. É de se notar que a quantidade de gráficos gerados cresce em razão quadrática à dimensão dos dados em si, pela natureza combinatória da abordagem. Mais especificamente, $\frac{m^2-m}{2}$ distintos, ou seja, ao analisar um dado de dimensão $m = 4$ são gerados 16 gráficos (6 distintos). Entretanto, tomando o dado multidimensional como imagens em tons de cinza de 16x16 pixels, podemos definir cada imagem como um ponto no $\mathbb{R}^{256}$, onde a intensidade corresponde à coordenada nesse espaço. Para visualizar esse dado, $m = 256$, são gerados 65.536 gráficos (32.640 distintos), constituindo uma limitação dessa abordagem (KEIM, 1997). Note que também possibilita responder como dimensões relacionam-se entre si, duas a duas, semelhante às coordenadas paralelas. Adicionalmente, possibilita que através da seleção de pontos em suas

projeções (*brushing*), seja evidenciado suas posições em todas os outros pares de dimensões (*linking*), como ilustrado na Figura 3b.

Figura 3 – SPLOM de dados de dimensão quatro, onde cada ponto é composto pelas medidas de largura e comprimento da pétala e da sépala de flores do gênero Íris. As diferentes cores refletem as três espécies diferentes visualizadas.



(a) SPLOM. (b) Operações de *linking* e *brushing*.

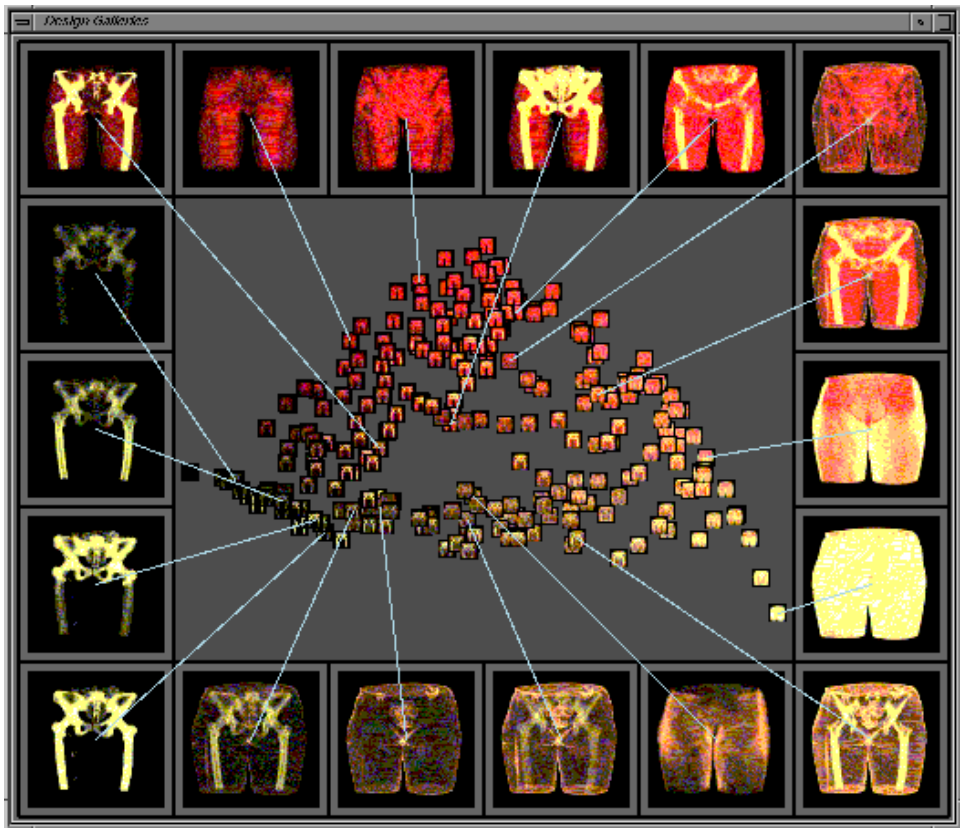
Fonte: Adaptada de Bostock (2012).

Uma alternativa bastante utilizada consiste em utilizar gráficos de dispersão após reduzir a dimensionalidade dos dados para o chamado *espaço visual*, isto é, $m \leq 3$, onde $m = 2$ é o mais utilizado. Sedlmair, Munzner e Tory (2013) investigam a utilização de $m = 3$ mostrando que o ganho de informação é restrito a poucos casos, não justificando seu uso em um caso geral, onde acaba havendo um aumento na dificuldade de análise.

Essa redução se dá através de técnicas de *projeção multidimensional*, que permitem projetar os dados de entrada em um espaço visual, isto é, de duas ou três dimensões de forma a preservar algum aspecto do dado multidimensional original tanto quanto possível. Dessa forma, a escolha de uma técnica em particular fica associada com o que deseja-se minimizar de perda durante a projeção: distância relativa entre os pontos (JOIA *et al.*, 2011) e topologia local (vizinhança) (PAULOVICH *et al.*, 2008) são dois exemplos bastante comuns. A utilização de gráficos de dispersão ajuda a fornecer uma informação de distâncias relativas entre pontos projetados no espaço visual e também permite interação em ambientes exploratórios (JEONG *et al.*, 2009; BROWN *et al.*, 2012). A metáfora equivale a observar na Figura 3 um único gráfico de dispersão. Na Figura 4, sua utilização é combinada (*linked-views*) com a metáfora de galerias com o objetivo de analisar um espaço de alta dimensão relativo às imagens médicas, onde pontos de controle parametrizam uma função de transferência, e as coordenadas de todos esses pontos, juntas, definem o dado dado multidimensional, com cada ponto projetado no espaço visual

correspondendo ao *rendering* obtido com a função de transferência associada (MARKS *et al.*, 1997). Utilizar esse tipo de metáfora é particularmente interessante em análise de espaços de parâmetros (BERGER *et al.*, 2011; TURKAY, 2013; RAUBER *et al.*, 2015).

Figura 4 – Explorando diferentes funções de transferência em um dado volumétrico da tomografia computadorizada de uma pelvis humana, com cada *rendering* associado a um ponto distinto em um gráfico de dispersão.



Fonte: Marks *et al.* (1997, Página 11).

A definição de dados multidimensionais é de certa forma algo geral, pois cada dimensão pode ser definida tanto atributos originais do objeto (por ex., medida física de plantas; intensidade do pixel na imagem) como atributos derivados (coeficientes da imagem em um espaço de características como o *Scale Invariant Feature Transform* - SIFT) (KE; SUKTHANKAR, 2004).

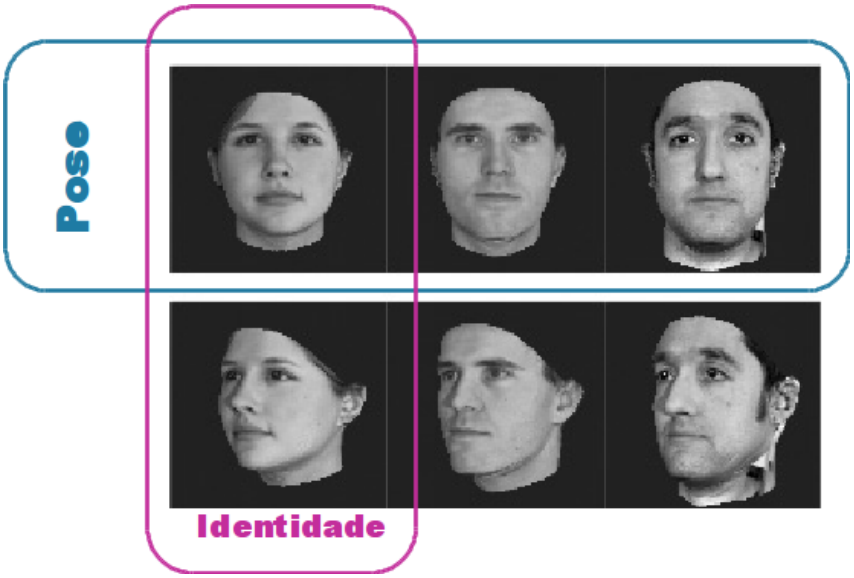
Essa definição por si só implica em uma grande quantidade possível de análise de um dado multidimensional que represente um mesmo objeto (por ex., imagem), que pode ser explorado por diversos aspectos. Isso implica que ao analisar uma projeção de um dado multidimensional deve-se levar em consideração o espaço original na qual ele foi definido, pois associa uma semântica ao processo. Nesse sentido, a visualização pode contribuir ao utilizar elementos visuais para simbolizar uma semântica, um comportamento que os dados possuam - isto é, através de uma *metáfora visual*.

Ao analisar a metáfora visual de gráficos de dispersão, por ex., a distância em que pontos distintos são desenhados deve refletir alguma medida de similaridade entre eles, de forma que pontos desenhados de forma mais próxima são pontos mais similares. Assim, analisar a diferença

entre similaridade percebida através de alguma metáfora visual e a similaridade calculada, de acordo com alguma métrica, é um objeto de estudo também em áreas ligadas à psicologia e percepção (BRINKE; SQUIRE; BIGELOW, 2004).

No exemplo da Figura 5 dois diferentes aspectos (pose e identidade) são ilustrados como possíveis medidas de similaridade a serem adotados. Vale salientar que esse aspecto semântico, que relaciona-se com o aprendizado de métricas de similaridade, apesar de importante, não será explorado de forma detalhada ao longo desse trabalho.

Figura 5 – Explorando medidas de similaridade em imagens de rostos de pessoas diferentes em poses diversas.



Fonte: Adaptada de Kulis (2013).

Coleção de dados multidimensionais

Ao definir um conjunto de dados multidimensionais pode-se utilizar o termo *coleção* quando dentro desse conjunto os dados de mesma natureza (por ex., imagem, documentos) podem ser separados em diferentes classes. Por exemplo, no contexto de imagens, um *conjunto* de dados multidimensionais conterá imagens de rostos de diferentes pessoas em uma mesma pose, ou seja, imagens da linha superior da Figura 5. Ao passo que uma *coleção* de dados multidimensionais poderia ser composta por imagens dos rostos de indivíduos diferentes em diferentes poses, ou seja, todas as imagens da Figura 5. O exemplo da Figura 4, por ex., não define uma coleção, mas sim ilustra um conjunto multidimensional de imagens.

Já no contexto de documentos científicos (ou seja, artigos), uma coleção pode ser definida através de documentos de diferentes tópicos publicados sob uma determinada grande área, por exemplo. Definir uma metáfora visual para tal tipos de coleção acaba sendo um grande desafio, pois envolve agregar tanto conceitos que podem ser numéricos (como a similaridade dentre documentos), quanto eventual relação semântica entre os documentos (citações entre

si); relevância individual de cada documento na coleção e elementos textuais (que envolvem milhares de palavras por documento).

1.2 Medidas de qualidade em projeção multidimensional

De uma forma geral, nem todo dado multidimensional tem uma representação visual significativa, como ilustrado na Figura 4, onde cada ponto está associado diretamente a uma imagem. Dessa forma, a representação por gráficos de dispersão ilustrará apenas pontos.

Na literatura, há alguns estudos voltados à avaliação quantitativa de forma a caracterizar gráficos de dispersão de acordo com diferentes padrões caracterizados por eles. Por exemplo, classificando-os através da análises de fatores de separação em *cluster* (SEDLMAIR *et al.*, 2012) e medidas gerais baseadas em grafos (WILKINSON; ANAND; GROSSMAN, 2005).

Por outro lado, como o processo envolve a avaliação de um usuário, e uma vez que a percepção humana é baseada na busca de padrões, a natureza dessa metáfora pode levar a um desentendimento sobre características dos dados originais, da projeção e o julgamento de qualidade da projeção realizado por um indivíduo. Na Figura 6 são ilustrados os gráficos de dispersão utilizando duas diferentes técnicas de projeção, com o mesmo conjunto de dados multidimensionais. Pode-se indagar: Qual delas é a correta? Qual é mais fiel aos dados?

Figura 6 – Projeção no espaço visual de um conjunto de imagens carimbos.

(a) Projeção utilizando PCA. (b) Projeção utilizando *Sammon mapping*.

Fonte: Elaborada pelo autor.

Dessa forma, um dos maiores desafios relacionados ao uso de gráficos de dispersão em projeções no espaço visual (*spatializations*) refere-se a sua efetividade, ainda que com perda de informação devido a projeção, em revelar características dos dados (por ex., *clusters*, tendências). Nesse contexto são definidas *métricas de qualidade*, que possibilitam analisar a fidelidade da projeção sob algum critério, como por ex. se a relação de vizinhança dos dados foi preservada,

possibilitando comparar diferentes projeções (BERTINI; TATU; KEIM, 2011; SEDLMAIR *et al.*, 2012; TATU, 2013).

Além disso, é importante notar que o espaço visual é definido através de duas dimensões, em geral. No contexto de SPLOMs, essas dimensões são definidas como subespaços alinhados aos eixos cartesianos do espaço original, com a projeção ortogonal sobre esse subespaço. Elas podem ser definidas também como um subespaço não alinhado aos eixos, como ocorre em técnicas como *Principal Component Analysis* (PCA), por exemplo, ou de abordagens não lineares, como o *Kernel Principal Component Analysis* (KPCA), onde um subespaço de dimensão dois é definido de forma sintética (SCHöLKOPF; SMOLA; MüLLER, 1998).

Centralidade

Uma forma de analisar a qualidade consiste em selecionar e visualizar pontos que tenham algum significado mais específico, de relevância (e.g., centralidade), dentro da base de dados.

Embora a média possa ser utilizada como uma medida com uma semântica de centralidade associada, seu uso pode levar a desentendimentos na interpretação dos dados uma vez que é fortemente influenciada por *outliers* e não simetrias na distribuição dos dados. Uma medida em particular ainda não explorada nesse contexto envolve a estimativa estatística de profundidade de dados (*data depth*), que está diretamente associada à noção de mediana dos dados. O uso da mediana como estimativa robusta de centralidade surge como uma alternativa interessante, uma vez que pode ser resistente a até 50% de *outliers* (REIMANN *et al.*, 2011). Estimadores estatísticos robustos têm se mostrado aplicáveis em outras áreas de pesquisa como processamento de imagem (SHAPIRA; AVIDAN; SHAMIR, 2009) e processamento geométrico (FLEISHMAN; COHEN-OR; SILVA, 2005).

Além disso, a noção de profundidade de dados está relacionada a análises estatísticas multivariadas e não-paramétricas. Nessa situação, nenhuma ou poucas suposições sobre a distribuição são utilizadas a priori (LIU; PARELIUS; SINGH, 1999; SERFLING, 2006).

1.3 Objetivos

Levando-se esses fatos em consideração, o presente trabalho objetiva

- Propor uma metáfora visual que possibilite visualizar, de forma combinada, diferentes aspectos de coleções de documentos científicos;
- Avaliar a utilização de medidas estatísticas robustas como uma medida de qualidade em projeções multidimensionais;
- Avaliar a utilização de medidas de centralidade em dados multidimensionais utilizando teoria de *Kernels*.

1.4 Contribuições

De forma resumida, levando em conta os objetivos propostos, as principais contribuições deste trabalho são:

1. Uma metáfora para visualização em multiescala de coleções de documentos científicos, que combina de forma compacta: a visualização sem intersecções de documentos por relevância, sumarização através de uma *word cloud* multisemente, noção de densidade por tópico e relação de citação entre diferentes documentos, possibilitando também a interação do usuário no processo;

2. Uma nova medida de qualidade para avaliação de técnicas de projeção multidimensional quanto à distorção introduzida por meio da utilização de funções de profundidade dados;

3. Uma metáfora visual que possibilite visualizar regiões quanto à sua centralidade e comparar a ação de técnicas de projeção multidimensional sobre elas;

4. Direcionar o processo de projeção multidimensional utilizando a informação de profundidade por diferentes estratégias de amostragem.

A contribuição (1) culminou na publicação:

- PAGLIOSA, P MARTINS, R.M., CEDRIM, D., F., PAIVA, A., MINGHIM, R., NONATO, L.G., *MIST: Multiscale Information and Summaries of Texts*, XXX Sibgrapi - Conference on Graphics, Patterns and Images, Arequipa - Peru 2013;

As contribuições (2,3,4) culminaram na publicação:

- CEDRIM, D., VAD, V., CASTELO, A., PAIVA, A., GRÖLLER, E., NONATO, L. G., *Depth functions as a Quality Measure and for Steering Multidimensional Projections*, Computer & Graphics, 2016, DOI: 10.1016/j.cag.2016.08.008

Para a compilação de todos os trabalhos desenvolvidos ao longo do doutorado, isto é, no período de 2011-2016, checar o Apêndice A.

1.5 Organização

No **Capítulo 2** é explorada uma metáfora visual que permite visualizar coleções de documentos científicos através de uma visualização combinada de: discos (modificação de gráficos de dispersão) e sumarização de conteúdo dos documentos (*word clouds*);

No **Capítulo 3** o aspecto de qualidade de uma projeção multidimensional é explorado através do conceito estatístico de profundidade de dados;

No **Capítulo 4** é analisado o aspecto qualidade através profundidade de dados no contexto de espaços de Hilbert através de funções *Kernel*;

E por fim, no **Capítulo 5** é feita uma discussão geral apontando as principais limitações identificadas e as propostas de trabalhos futuros a serem desenvolvidos.

CAPÍTULO

2

VISUALIZAÇÃO DE COLEÇÃO DE DOCUMENTOS CIENTÍFICOS

2.1 Aspectos gerais

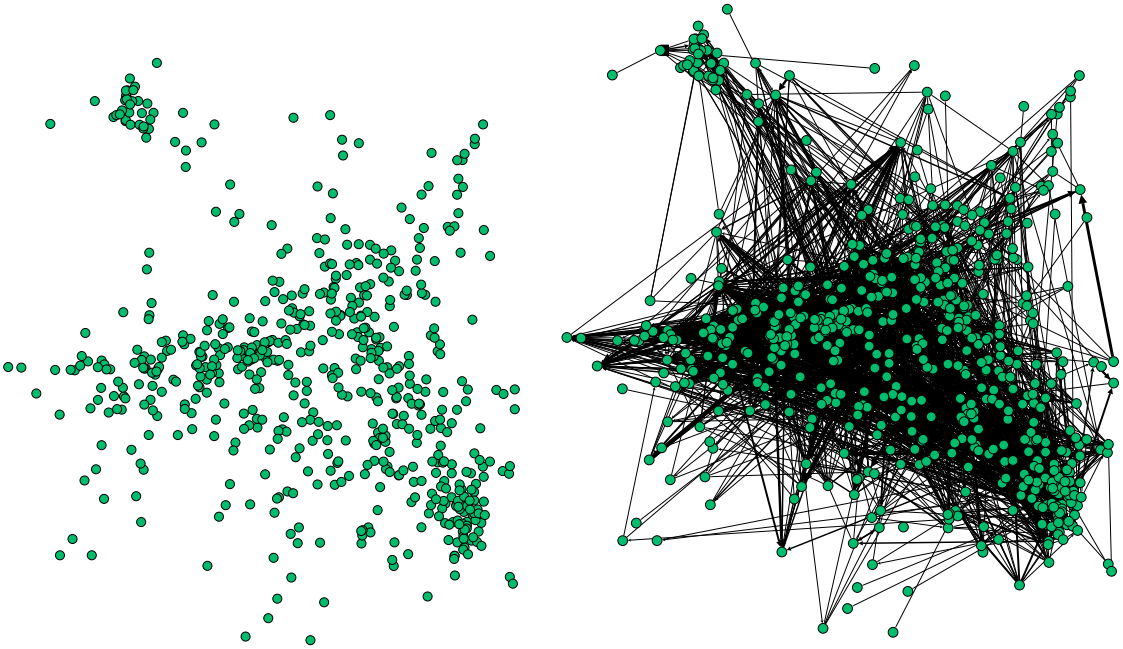
No cenário atual, de um crescimento sem precedentes de dados textuais, um grande número de aplicações pode ser afetado por meio de uma exploração de coleções de documentos, visando a identificação e extração de informações úteis. Dentro desse cenário, documentos de textos podem ser mais simples, no sentido de conter apenas informação textual e eventualmente imagens, independentes entre si dentro de uma coleção. Podem também conter informações adicionais que os relacionem entre si, como no contexto científico acontece com artigos, que possuem ligações com uma semântica associada (ou seja, citação entre artigos).

Uma possível metáfora para visualização de uma coleção de documentos com essas características pode ser observada na Figura 7. Na Figura 7a é ilustrada uma visualização comum por gráfico de dispersão, que deixa de transmitir visualmente uma informação relevante da natureza desse tipo de coleção, às citações entre os documentos. Por outro lado, quando considera-se transmitir tal informação, como ilustrado na Figura 7b, pode haver dificuldade em analisar a relação de citações entre os nós devido à eventual sobreposição de informações.

Assim, a utilidade de um dado método de visualização de uma coleção de documentos depende largamente do quão eficientemente a metáfora visual adotada sintetiza e modifica a informação a ser extraída por meio da visualização. Por exemplo, nuvens de palavras (*word clouds*) podem ser consideradas efetivas em aplicações onde o objetivo é prover visualmente um resumo do conteúdo de documentos. Por outro lado, métodos baseados em força são mais apropriados em situações que demandam a identificação e manipulação de documentos específicos ou grupos de documentos relacionados.

Na Figura 8 é ilustrado o resultado de um algoritmo simples de sumarização utilizando

Figura 7 – Exemplo de visualizações de coleção de documentos científicos.



(a) (b)

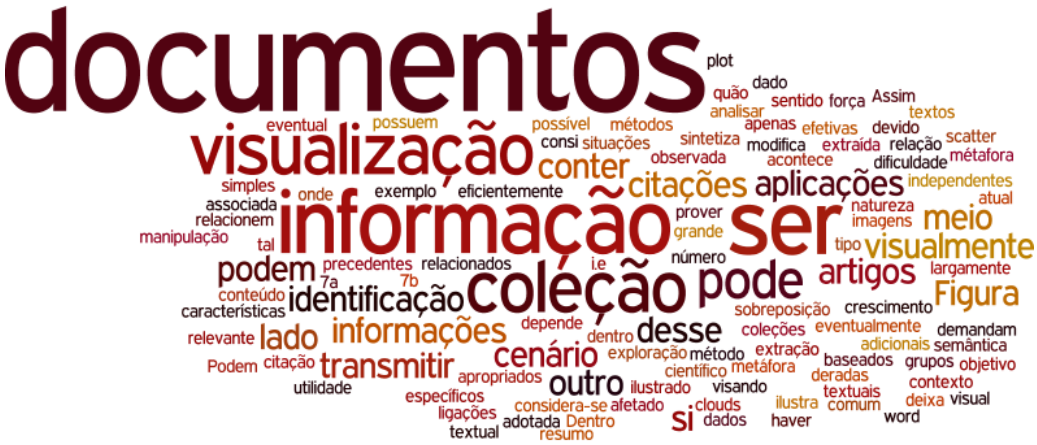
Fonte: Elaborada pelo autor.

nuvens de palavras (VIéGAS; WATTENBERG; FEINBERG, 2009). Ela sumariza o texto dos quatro primeiros parágrafos desta Seção, onde nitidamente entre os termos de maior destaque há: "visualização", "informação"e "documentos", que refletem uma boa sumarização do conteúdo dos dois parágrafos.

Com o objetivo de propiciar um conjunto rico de informações em uma única visualização, diversas metodologias sugerem a combinação de múltiplas metáforas em um *layout* unificado. Embora múltiplas metáforas favoreçam a apresentação simultânea de informações de natureza distintas, são poucas as abordagens que de fato são efetivas no processo de *layout* compostos, que forneçam um resultado que carregue significado e que ao mesmo tempo evite distrações e poluição visual. Em particular, a combinação de *layouts* de pontos dinâmicos de base de textos e resumos baseados em conteúdo (por ex., *word clouds*) é um desafio ainda não propriamente esclarecido. Os poucos métodos que propõem uma integração dessas duas metáforas de visualização são ainda deficientes em termos de: a) qualidade do *layout* resultante; b) limitados em relação a interatividade; c) não são diretamente escaláveis.

É neste contexto que surge a metodologia *Multiscale Information and Summaries of Texts* (MIST) de forma a fornecer uma *word cloud* capaz de bem explicitar documentos e grupos de

Figura 8 – Exemplo de nuvens de palavras para sumarização de texto, utilizando a técnica de Viegas et al. (VIéGAS; WATTENBERG; FEINBERG, 2009).



Fonte: Elaborada pelo autor.

documentos similares em um plano de visualização, além de permitir uma manipulação interativa por parte do usuário.

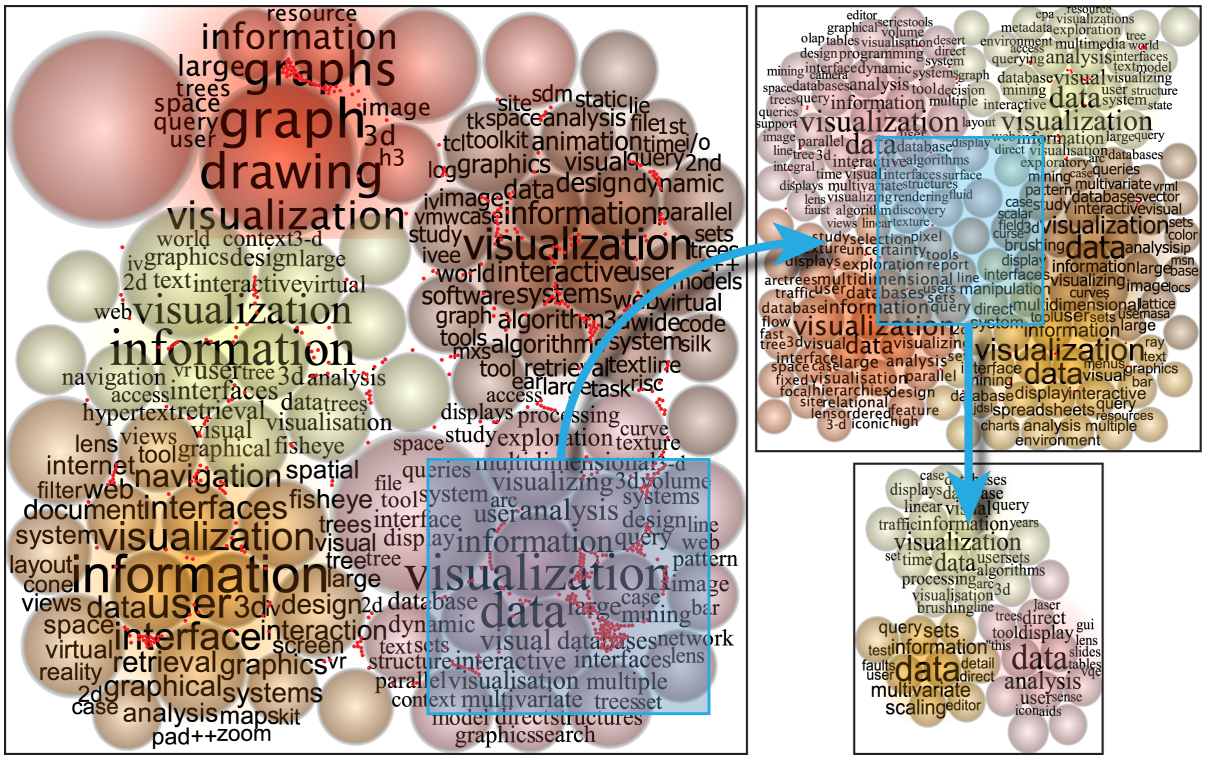
A metodologia relaciona uma dada *word cloud* a cada grupo de documentos similares via um *layout* gerado através de uma simulação de corpo rígido. Essas *word clouds* são harmoniosamente dispostas em um espaço visual utilizando um esquema multi sementes. Uma simulação de corpo rígido organiza discos - que representam aqui documentos - de forma a garantir que instâncias similares sejam dispostas próximas umas as outras, como ilustrado na Figura 9.

A formulação também leva em consideração a relevância de cada documento na coleção ao atribuir tamanhos de disco diferenciados em acordo com a importância de cada documento. Um *layout* claro e intuitivo é mantido uma vez que apenas documentos relevantes são retratados como discos. Documentos fora dos limites de relevância podem ser simplesmente removidos da visualização ou podem ser representados por pequenos *glyphs* como forma de se preservar a percepção de densidade.

Instâncias fora da janela de visualização não são utilizadas durante uma simulação de corpo rígido como forma de acelerar o processo. No entanto, essas instâncias deslocam-se em conformidade com seus vizinhos de forma a serem recuperadas, se necessário. Essa estratégia pode auxiliar na escalabilidade da simulação de corpo rígido no espaço visual.

A abordagem é altamente interativa, flexível e intuitiva. Os discos podem ser arrastados ao redor do espaço visual objetivando a geração de novos *layouts*. O usuário pode ainda direcionar a visualização a regiões específicas do espaço visual e com isso explorar subconjuntos de

Figura 9 – MIST representando um *layout* combinando projeção multidimensional, simulação de corpo rígido, sumarização via *word clouds* e multiescala via *ranking* e operação de *zoom*.



Fonte: Elaborada pelo autor.

documentos. Vale-se ressaltar que durante a navegação, o *layout* é dinamicamente atualizado de forma a garantir que não existam estruturas e conteúdo escondidos.

As principais contribuições da metologia são:

- O uso de uma *engine* de simulação de corpo rígido como forma de organizar documentos em um espaço visual de forma a preservar a estrutura de vizinhança e evitar a sobreposição de documentos, representados por nós de diferentes tamanhos;
- Propor uma estratégia original que combina simulação de corpo rígido e *word clouds*, no sentido de possibilitar o reconhecimento de similaridades entre diferentes documentos e seus respectivos conteúdos de uma forma unificada;
- Uma nova abordagem para construção de *word clouds* partindo-se de um conjunto de múltiplas sementes dispostas abaixo dos *clusters*.

Resultados aqui apresentados suportam a afirmação de que a combinação harmoniosa de metáforas juntamente com recursos gráficos e interativos enriquecem o sistema com um conjunto não usualmente disponíveis por outras estratégias de visualização Avanços no estado da arte na área de análise de documentos são observados ao se unificar essas características em uma única apresentação.

2.2 Trabalhos relacionados

A visualização de coleção de documentos é uma área de pesquisa significativamente produtiva, fazendo com que muitos dos métodos existentes variem significativamente em suas bases matemática e computacional. Com o intuito de contextualizar a metodologia MIST, uma breve descrição de métodos de visualização ligados a coleções de documentos é organizada de acordo com sua metáfora visual. Evidencia-se principalmente propriedades gerais e limitações de cada caso e evita-se um detalhamento excessivo. Ao leitor interessado, uma descrição mais detalhada pode ser encontrada em um *survey* (ALENCAR; OLIVEIRA; PAULOVICH, 2012).

*Nuvens de palavras (*Word clouds*)*

O trabalho de Kuo *et al.* (2007) pode ser considerado com um dos precursores no uso de palavras-chaves como recurso de visualização ao propor uma organização linha a linha de palavras de acordo com relevância. A limitação relacionada a grande quantidade de espaços em branco apresentada pela técnica de Kuo foi explorada por Kaser e Lemire (2007) e Wordle (VIé-GAS; WATTENBERG; FEINBERG, 2009). Ambos os trabalhos tem como produto *layouts* mais agradáveis e visualmente atraentes. A falta de uma relação semântica entre palavras em uma *cloud* foi explorada por ManiWordle (KOH *et al.*, 2010) através de um mecanismo interativo assistido e por Cui *et al.* (2010) por meio de um método de força o qual também permite uma visualização da estrutura temporal dos documentos.

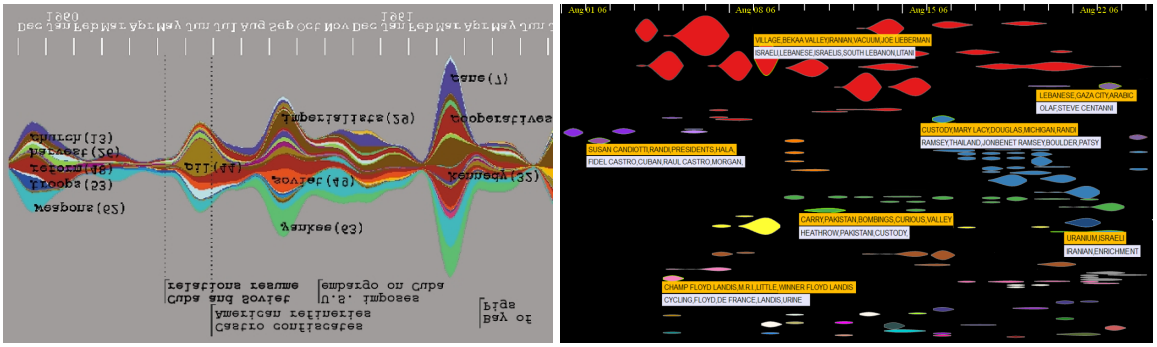
Wu *et al.* (2011) apresentaram uma metodologia que primeiramente calcula relacionamentos semânticos e posteriormente utiliza projeção multidimensional para organizar palavras-chaves no espaço visual. Uma abordagem utilizando a relevância de palavras através do vetor de Fiedler foi utilizado para ordenação semântica em Paulovich *et al.* (2012) e apresentou resultados bastante efetivos, ilustrado na Figura 25a. Outros métodos como SparkClouds (LEE *et al.*, 2010) e Parallel Tag Clouds (COLLINS; VIEGAS; WATTENBERG, 2009) expandem a *word cloud* por meio da aplicação de recursos extra visuais como *sparklines* e coordenadas paralelas, como forma de melhor transmitir o sumário de conteúdo dos documentos. Embora bastante efetivos no processo de explicitar informações essenciais contidas em uma coleção de documentos, o paradigma de *word clouds* por si só não identifica associações de palavras em um determinado documento ou conjunto de documentos, nem tampouco permite uma análise de similaridade entre documentos.

*Metáfora de rio (*River Metaphor*)*

Metáforas de rio são conhecidas como um mecanismo efetivo na visualização de mudanças temáticas temporais em coleções de documentos. O conceito, introduzido no sistema ThemeRiver (HAVRE *et al.*, 2002), ilustrado na Figura 10a, vem sendo melhorado por meio do uso de sofisticados mecanismos no processo de derivação de *time-sensing topics* (LIU *et*

al., 2012) e visualização de camadas capazes de descrever o nascimento, morte e divisão de eventos (CUI *et al.*, 2011). EventRiver (LUO *et al.*, 2012) utiliza um esquema de clusterização para agrupar documentos que são similares em conteúdo e próximos em tempo. No esquema, a espessura de uma bolha representa o número de documentos e o período de duração de um evento, ilustrado na Figura 10b. *History Flow* (VIéGAS; WATTENBERG; DAVE, 2004) pode ser também observado como uma metáfora *river-based* projetada para visualizar edições de um documento (ou coleção de documentos como por exemplo a Wikipédia) constituído por diferentes autores, de forma a enfatizar quais partes persistem ao longo do tempo.

Figura 10 – Visualização de duas abordagens que utilizam a metáfora de rio para sumarizar importância em coleções de documentos, utilizando adicionalmente o aspecto de evolução temporal.



(a) Theme river.

Fonte: Havre *et al.* (2002).

(b) Event river.

Fonte: Luo *et al.* (2012).

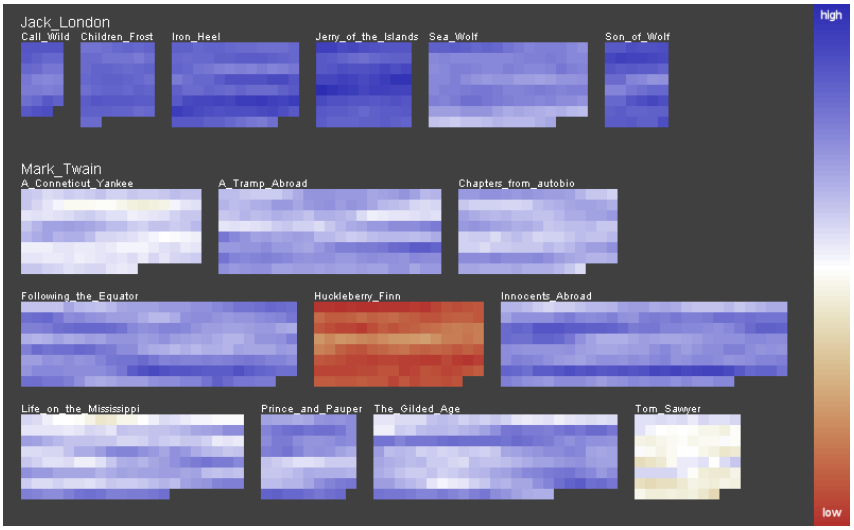
Metáforas de rio fornecem uma visualização agradável e intuitiva ao explicitar o comportamento temporal de uma coleção de documentos mas, de forma similar a *word cloud*, não permitem a identificação imediata de documentos específicos, sua relevância dentro do conjunto ou sua contribuição em relação a um determinado tópico. Além disso, a interação da técnica com o *river layout* no sentido de realizar modificações na perspectiva do usuário frente aos conjuntos de dados não é factível.

Linguística

Métodos que constroem visualizações baseadas em estruturas linguísticas semanticamente definidas também tem sido propostas na literatura. *Word Tree* (WATTENBERG; VIéGAS, 2008), por exemplo, utiliza um *layout* de árvore para visualizar a ocorrência de termos junto a frases próximas, nas quais são rearranjadas como ramos descendentes da árvore. *Phrase Nets* (HAM; WATTENBERG; VIEGAS, 2009) emprega um *layout* baseado em grafo em que cada nó representa um subconjunto de palavras e cada aresta corresponde a uma relação lexical ou semântica entre palavras. O tamanho de fonte e a espessura de borda são utilizados para visualmente marcar atributos como o número de ocorrências de um conjunto de palavras e seus relacionamentos. Uma análise linguística mais sofisticada é aplicada por DocuBurst (COLLINS; CARPENDALE; PENN, 2009), no qual faz uso de uma base de dados lexical eletrônica e de

uma *radial space-filling tree layout* para visualizar o conteúdo de um documento lexicalmente. Keim e Oelke (2007) propuseram um método que emprega regras semânticas para segmentar um documento em blocos e *function words* (i.e. preposições, pronomes, conjunções) e mapear esses blocos em vetores de características. O componente principal de cada vetor é utilizado para colorir o bloco propiciando ao documento uma identidade visual. Na Figura 11 é ilustrado esse trabalho, e é interessante notar que a construção permite uma forma de comparação de autoria através da análise do estilo de escrita de cada autor. No exemplo ilustrado, são visualizadas algumas obras de Jack London e Mark Twain. Particularmente, é interessante notar que a obra *Huckleberry Finn* deste apresenta um estilo de escrita diferente das demais do mesmo autor.

Figura 11 – Comparação de documentos utilizando a primeira componente principal da matriz de frequências de *function words*.



Fonte: Keim e Oelke (2007).

Diferentemente de outros métodos baseados em linguística anteriormente descritos, o método de Keim e Oelke (2007) permite a identificação e comparação de documentos específicos em um conjunto de dados, embora comprometa a legibilidade do conteúdo.

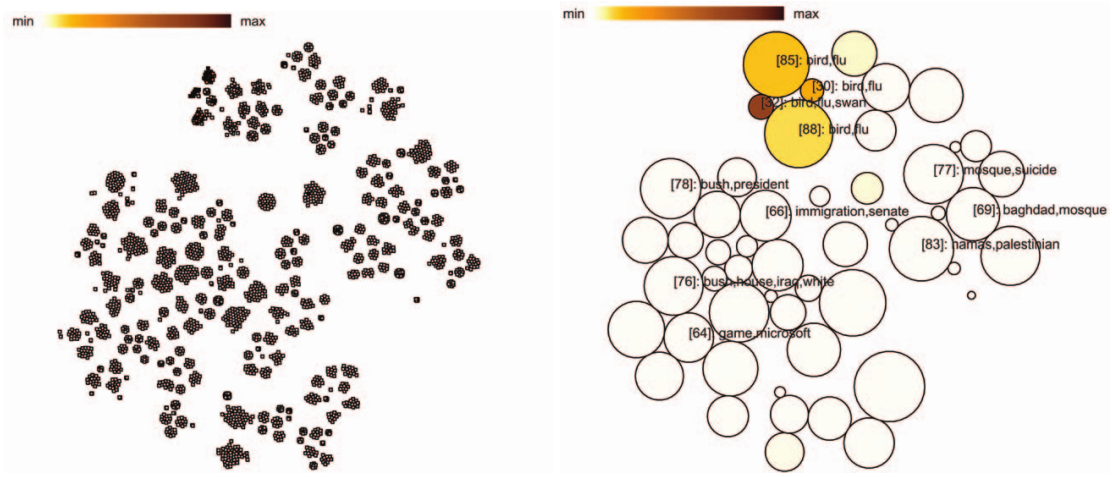
Hierárquico

Técnicas baseadas em estruturas hierárquicas visando a exploração e navegação de conteúdo com diferentes graus de detalhe são também encontrados na literatura. A técnica Topic Island (MILLER *et al.*, 1998), por exemplo, constrói uma hierarquia por meio da aplicação de uma análise de *wavelet* de um sinal extraído de palavras do documento. Essa hierarquia possibilita uma visualização de mudanças temáticas da coleção de documentos, bem como de suas partes mais importantes.

InfoSky (ANDREWS *et al.*, 2002) visualiza documentos hierarquicamente organizados por meio de uma subdivisão do espaço visual utilizando um diagrama de Coronoide recursivo. A navegação pela estrutura fica possibilitada por meio de um mecanismo do tipo *telescope-like*

zoom. HiPP (PAULOVICH; MINGHIM, 2008) faz uso de uma *cluster tree* de forma a organizar documentos hierarquicamente de acordo com suas similaridades, provendo a visualização através da projeção nos nós da árvore, na Figura 12 essa abordagem é ilustrada. Mao, Dillon e Lebanon (2007) visualizam documentos através de curvas construídas por meio de uma generalização de n-grams e médias locais e constrói a hierarquia por meio da modificação do suporte dos *kernels* utilizados na média computacional.

Figura 12 – Visualização da abordagem hierárquica HiPP para a coleção de dados de *RSS feeds* de várias fontes.



(a) Visualização antes da clusterização hierárquica.

Fonte: Paulovich e Minghim (2008).

(b) Visualização após a clusterização hierárquica.

Fonte: Paulovich e Minghim (2008).

Embora efetiva na construção de resumos visuais e na identificação de estruturas nos documentos, técnicas hierárquicas não são efetivas na associação entre conteúdo e documentos quando a hierarquia é realizada sobre tópicos. Além disso, a visualização simultânea da estrutura hierárquica e da importância de cada documento não é uma tarefa trivial.

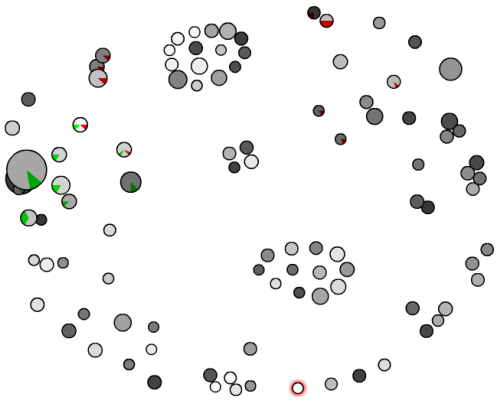
*Direcionado por Força (*Force Directed*)*

Técnicas do tipo *Force directed* constroem visualizações por meio da minimização de um funcional definido por similaridade de texto entre pares (*pairwise text similarity*). FacetAtlas (CAO *et al.*, 2010) emprega um *layout* de grafo baseado em força para estabelecer *clusters* de nós e enriquecer a visualização utilizando estimadores de densidade e representações de dados por múltiplas faces.

TopicNets (GRETARSSON *et al.*, 2012) computa a dissimilaridade entre tópicos extraídos de uma coleção de documentos e aplica uma técnica de projeção no processo de organização desses tópicos na janela visual disponível. Na sequência, um mecanismo do tipo *force directed* convencional é aplicado visando a organização de cada documento ao redor dos respectivos tópicos. No trabalho STREAMIT (ALSAKRAN *et al.*, 2012) é utilizada a similaridade entre documentos para definir forças e a lei de Newton para atualizar a posição de cada nó no espaço

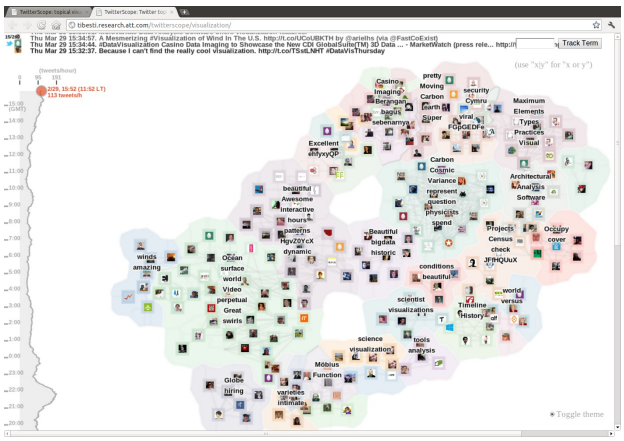
visual. Documentos podem ser adicionados dinamicamente no sistema, possibilitando assim a visualização de um fluxo de documentos, ilustrado na Figura 13a. O fluxo de dados é também considerado em TwitterScope (GANSNER; HU; NORTH, 2012), onde há uma conversão entre o grafo de similaridade e um *layout* de mapas geográficos, e a aplicação de um esquema de forças no sentido de remover interseções entre nós. Como ilustrado na Figura 13b, essa abordagem não permite a análise de relevância individual de cada elemento, mas de regiões compostas por vários elementos, contrário ao que analisamos no presente trabalho.

Figura 13 – Visualização de abordagens direcionadas por força.



(a) STREAMIT com coleção de documentos da *National Science Foundation* - NSF.

Fonte: Alsakran *et al.* (2012).



(b) TwitterScope com mensagens do Twitter com o termo "visualização"para um horário fixado.

Fonte: Gansner, Hu e North (2012).

De forma geral, o tratamento empregado ao problema de interseção tem sido tópico de discussão em uma série de trabalhos (SPRITZER; Dal Sasso Freitas, 2012; FRUCHTERMAN; REINGOLD, 1991; GEIPEL, 2007; GIBSON; FAITH; VICKERS, 2012; HU, 2005), muito embora não se tenha ainda garantia de preservação da estrutura de vizinhança durante a simulação.

Métodos baseados em força permitem a identificação de um documento particular e sua relação com a vizinhança. No entanto, tais métodos não são efetivos em prover um resumo visual do conteúdo dos documentos. A técnica de *wordification* proposta por Paulovich *et al.* (2012) supre essa limitação ao combinar um algoritmo de *force directed*, uma projeção multidimensional e uma *word cloud*, permitindo uma visualização onde documentos similares ficam dispostos próximos uns aos outros no espaço visual. Sumários de conteúdo são também apresentados por meio de uma *word cloud* construída através dos *clusters* de documentos. Em contrapartida, a abordagem apresentada pelos autores não permite a interação com o *layout* de forma a gerar arranjos por meio da inserção de informação como grau de importância de cada documento.

Outras metáforas

Chuang *et al.* (2012) propôs o Stanford Dissertation Browser, disponibilizando um conjunto de recursos de visualização úteis à investigação do impacto de pesquisas interdisciplinares entre departamentos da Universidade de Stanford. Técnicas de projeção baseadas em PCA e *layouts* radiais são utilizados para investigar visualmente o conjunto de ideias compartilhadas e o processo de colaboração.

Document Cards (STROBELT *et al.*, 2009) apresenta uma visão geral de uma coleção de documentos ou de documentos por meio da adoção de uma análise racional de *top trumps game cards*, caracterizadas pelo uso de imagens expressivas e fatos para ressaltar termos chaves e imagens relevantes extraídas de um documento. As estratégias de Chuang e de *Document Cards* são adequadas no sentido de fornecer uma visualização compacta de uma grande coleção de documentos. No entanto, ambas as técnicas falham no processo de fornecer informações sobre relacionamentos entre documentos.

As metodologias já existentes não são projetadas para associar mecanismos interativos e dinâmicos, no sentido de se construir um sumário visual de uma dada coleção de documentos, possibilitar a visualização de similaridades entre documentos, a importância de cada elemento na coleção e seus relacionamentos como, por exemplo, *links* e citações. Desconhece-se atualmente técnica disponível capaz de integrar todas essas operações. Como consequência, técnicas usualmente desenvolvidas forçam o usuário a utilizar diferentes tipos de visualização, que podem não estar necessariamente conectadas.

A técnica aqui descrita engloba um conjunto de características que suprem a necessidade de funcionalidade simultânea, tornando possível a visão geral do conteúdo de um documento em uma coleção, além do estabelecimento de correspondências entre documentos e palavras. Adicionalmente, o usuário pode rearranjar o *layout* dinamicamente como forma de explorar um subconjunto de documentos e sua vizinhança, possibilitando a recriação das sumarizações baseadas no novo posicionamento.

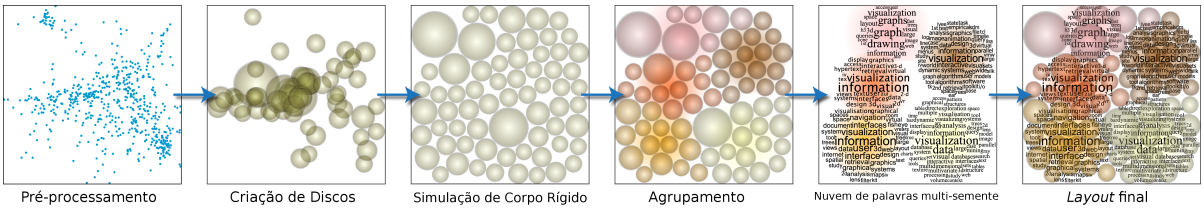
2.3 A técnica MIST

A técnica MIST é composta por cinco etapas: pré-processamento, criação de discos, simulação de corpo-rígido (*ridig-body simulation*), agrupamento e geração de uma nuvem de palavras (*word cloud*), como ilustrado na Figura 14.

A primeira etapa é o processo de extração de palavras-chave para geração do documento científico e da nuvem de palavras na última etapa do *pipeline*. Essas palavras-chave são utilizadas no processo de identificação de similaridade entre documentos. Essa similaridade é utilizada como argumento de entrada para um método de projeção multidimensional que mapeia documentos para um espaço visual bidimensional.

Cada ponto projetado corresponde a um documento, onde a importância de cada documento na coleção pode ser obtida de duas formas: informada pelo usuário previamente ou baseada em um grafo de conectividade entre cada documentos. Essa informação é utilizada na segunda etapa para criação dos discos que possam codificar visualmente essa informação.

Figura 14 – Etapas do *pipeline* do MIST.



Fonte: Elaborada pelo autor.

Na terceira etapa da MIST, uma *engine* de corpo-rígido organiza o conjunto de documentos, que são representados como discos, de forma que o tamanho de cada disco é considerado a fim de evitar possíveis interseções e melhor preservar a estrutura de vizinhança provida pela projeção multidimensional.

Na quarta etapa os documentos são agrupados de acordo com suas vizinhanças no espaço visual, o que definirá quais termos serão utilizados na criação da nuvem de palavras.

Na última etapa do *pipeline*, as nuvens de palavras são geradas e harmoniosamente dispostas em conjunto, de forma a compor o *layout* final. Detalhes técnicos de cada etapa são descritos na sequência.

Pré-processamento

Um documento científico possui uma estrutura bem definida, a saber: título, resumo, palavras-chave, corpo do documento (eventualmente contendo imagens e tabelas), citações a outros artigos como referências. No entanto, para fins de simplificação numérica, apenas os elementos textuais são considerados no contexto desse trabalho. Além disso, a informação de citação entre artigos define naturalmente um grafo direcionado, onde cada documento é um nó do grafo, cuja aresta é dada pela citação.

Consideramos também que todos os documentos analisados contêm elementos textuais em um mesmo idioma, o que será fundamental nessa etapa.

Durante o estágio de pré-processamento três tarefas são executadas: Extração de palavras-chave, representação de um documento cientifico, definição de medida de relevância de cada documento da coleção.

Tanto a construção da nuvem de palavras como a geração do espaço vetorial de representação para a projeção multidimensional dependem do conjunto de palavras-chaves extraído da coleção de documentos.

Extração de palavras-chave e Definição de um documento científico

Em todos as partes descritas da estrutura de um documento há elementos textuais envolvidos, definindo, dessa forma, a necessidade de efetuar um processamento textual para possibilitar a caracterização do documento em si. Dentro desse tópico de processamento textual há várias abordagens possíveis para extrair informações semântica de um documento de texto, dentre as quais as mais comuns são o modelo vetorial (*bag-of-words*) e modelo probabilístico *Latent Dirichlet Allocation* (LDA) (SALTON, 1991). Nesse trabalho decidimos utilizar o modelo vetorial como um ponto de partida, por sua simplicidade e bons resultados, que será descrito na etapa de definição do documento. Cabe notar quer *LDA* também poderia ter sido utilizado sem perda de generalidade da abordagem.

Utilizar o modelo vetorial possibilita definir o documento a partir de uma matriz onde cada linha corresponde a um documento e cada coluna contém um valor associado à frequência do termo naquele documento - *Term Frequency* (TF). Mais precisamente, utiliza-se a medida *Term Frequency-Inverse Document Frequency* (TF-IDF) como a entrada de cada elemento (i, j) da matriz, de forma que, seja $\mathbf{d}_i$ um i-ésimo documento da coleção de n documentos, $\mathbf{t}_j$ é o vetor (coluna) formado pela frequência do j-ésimo termo em cada documento, a medida é definida como:

$$w_{ij} = freq(\mathbf{d}_i, \mathbf{t}_j)\ log\frac{n}{d(\mathbf{t}_j)}, \tag{2.1}$$

onde $d(\mathbf{t}_j)$ corresponde ao número de documentos da coleção onde o j-ésimo termo ocorre.

A intuição dessa ponderação pelo fator *Inverse Document Frequency* (IDF) é para definir um maior peso para termos que apareçam em poucos documentos, aumentando assim a possibilidade de distinção de documentos semelhantes.

Adicionalmente, o corte de Luhn's foi efetuado (LUHN, 1958), de forma que retira todas as entradas que tenham frequência inferior a um limite inferior de corte. Ao longo dos exemplos investigados no trabalho esse limite foi definido como dez. Isso implica que a Equação 2.1 é bem definida para os valores calculados possíveis, já que com isso $d(\mathbf{t}_j) > 0, \forall j = 1, \ldots, n$.

Dessa forma, a matriz que representa a coleção dos documentos terá a seguinte forma:

$$\mathbf{M} = \begin{array}{c} \begin{array}{cccc} \mathbf{t}_1 & \mathbf{t}_2 & & \mathbf{t}_m \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1m} \\ w_{21} & w_{22} & \ldots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \ldots & w_{nm} \end{pmatrix} \end{array} \begin{array}{c} \\ \mathbf{d}_1 \\ \mathbf{d}_2 \\ \\ \mathbf{d}_n \end{array}, \tag{2.2}$$

onde w_{ij} é dado pela Equação 2.1.

Por construção, cada linha da matriz $\mathbf{M}$ representa um documento $\mathbf{d}_i$, onde as colunas

não-nulas representam os termos $\mathbf{t}_j$ contidos no documento $\mathbf{d}_i$. Vale observar que m representa o total de termos únicos da coleção, que eventualmente será bem maior que a quantidade de termos de qualquer documento individualmente. Assim, a matriz $\mathbf{M}$ construída diretamente do texto não-processado (*raw data*), é esparsa e de alta-dimensão. Algumas medidas podem ser adotadas com o objetivo de reduzir os efeitos do mal da dimensionalidade (*curse of dimensionality*), o que implica em melhorar a semântica extraída no processo.

A primeira delas é remover palavras com baixo teor semântico, por exemplo, devido a aparecem muito frequentemente em uma língua, podendo envolver: preposições, artigos, verbos, advérbios. Tais palavras são chamadas de *stopwords*. Note que esse processo pode variar de acordo com a língua utilizada. Basicamente as *stopwords* são compiladas em uma lista, já que em uma língua específica, e sua remoção dos termos dos documentos é baseada nessa lista. Na Figura 15 algumas dessas palavras são ilustradas. Cabe citar que um possível problema com esse processo é a possibilidade de omitir alguns termos técnicos de uma área, como por exemplo a exclusão do termo "*least*", cujo termo "*least squares*" possui grande importância em diversas áreas. Isso não constitui uma limitação na abordagem, apenas deve ser levado em consideração nessa etapa de construção das *stopwords*.

Figura 15 – Exemplo de nuvem de palavras para sumarização da lista de *stopwords*, utilizando a técnica de Viegas et al. (VIéGAS; WATTENBERG; FEINBERG, 2009).

Fonte: Elaborada pelo autor.

A segunda tarefa consiste em extrair os radicais das palavras, processo chamado de *stemming*, de forma que "*effective*"e "*effectiveness*" serão consideradas a mesma palavra por terem o mesmo radical. Nessa etapa utilizamos o algoritmo de Porter (1980). Note que ao efetuar essa operação a dimensão da matriz $\mathbf{M}$ é reduzida ainda mais.

A extração de palavras chave para definição dos documentos científicos considerou os

seguintes elementos textuais de cada documento da coleção: titulo do artigo, resumo e palavras-chave. No exemplo do *dataset* IV04 (IEEE Information Visualization) por ex., descrito na Seção 2.4, a matriz representada pela Equação 2.2 inicialmente tinha tamanho (614×6030), onde após essas operações foi reduzida para (614×582), apresentando uma redução de dimensionalidade da ordem de aproximadamente 90%.

É importante notar que a frequência das palavras é contabilizada de forma individual, reduzindo a representatividade semântica de palavras com mais de um termo contíguo, por exemplo "*data mining*", que terá contabilizado os termos independentes "*data*"e "*mining*". A abordagem por TF-IDF constitui uma heurística simples mas bastante utilizada em processamento de linguagem natural. Entretanto, outras abordagens mais sofisticadas como a exploração de *n-grams* podem ser considerada nessa etapa.

Projeção Multidimensional

A simulação de corpo rígido assume como condição inicial um conjunto de discos no espaço visual. O centro de cada disco é obtido pelo uso de um mapeamento de documentos definidos como vetores em um espaço de características para o espaço visual.

Na MIST, a projeção é efetuada utilizando a *Least Square Projection* (LSP) (PAULOVICH *et al.*, 2008) tendo distâncias euclidianas como medida de dissimilaridade utilizada, já que a efetividade e precisão da aplicação da LSP, no processo de projeção de dados textuais, já é bem estabelecido.

LSP é uma técnica de projeção multidimensional semi-supervisionada, ou seja, possibilita a utilização de pontos de controle para guiar o processo de projeção. Por construção ela visa preservar a relação de vizinhança tanto quanto possível. Para tal, considera os pontos no espaço original como combinações convexas de seus vizinhos e as coordenadas dos pontos de controle no espaço visual. Com isso, determina as posições dos pontos projetados tentando preservar esses coeficientes da combinação, bem como as posições relativas aos pontos de controle tanto quanto possível, em uma abordagem de mínimos quadrados. Mais formalmente, seja $\tilde{\mathbf{d}}_i$ a projeção no espaço visual de um documento $\mathbf{d}_i$, isto é, um ponto em $\mathbb{R}^m$. Escrevendo-o como combinação convexa de sua vizinhança V_i no espaço original temos:

$$\tilde{\mathbf{d}}_i - \sum_{\mathbf{d}_j \in V_i} \alpha_{ij} \tilde{\mathbf{d}}_j = 0, \tag{2.3}$$

com $0 \leq \alpha_{ij} \leq 1; \sum \alpha_{ij} = 1$, os coeficientes da combinação em $\mathbb{R}^m$.

Para incorporar as restrições dadas pela Equação 2.3 e pelos pontos de controle o seguinte sistema linear representado na Equação 2.4 deve ser resolvido para cada coordenada do espaço visual. Assim:

$$\begin{pmatrix} \mathbf{L} \\ \mathbf{C} \end{pmatrix} \mathbf{x} = \mathbf{b}, \tag{2.4}$$

onde $\mathbf{L}$ é a matriz de ordem n com os coeficientes de todos os pontos na forma:

$$\mathbf{L} = \begin{cases} 1 & i = j \\ -\alpha_{ij} & \mathbf{d}_j \in V_i \\ 0 & \text{caso contrário} \end{cases}, \tag{2.5}$$

$\mathbf{C}$ a matriz $n_{cp} \times n$ na forma, cujas linhas representam os n_{cp} pontos de controle cujos índices são representados pelas entradas não-nulas na forma:

$$c_{ij} = \begin{cases} 1 & \text{na j-ésima coluna se } \mathbf{d}_j \text{ é i-ésimo ponto de controle} \\ & \text{caso contrário} \end{cases}, \tag{2.6}$$

$\mathbf{b}$ uma matriz $(n+n_{cp}) \times 1$ de zeros relativos à restrição da Equação 2.3 e contendo uma das l coordenadas de cada k-ésimo ponto de controle $\tilde{\mathbf{d}}_k = (d_k^1, \ldots, d_k^l)$ no espaço visual, isto é:

$$b_i = \begin{cases} 0 & i \leq n \\ \tilde{d}_{i-n}^l & n < i \leq n+n_{cp} \end{cases}. \tag{2.7}$$

Como estamos considerando o espaço visual bidimensional, os pontos de controle são na forma $\tilde{\mathbf{d}}_k = (d_k^1, d_k^2)$, e é resolvido um sistema linear na forma da Equação 2.4 para determinar as coordenadas dos pontos na primeira dimensão da projeção e outro para as coordenadas da segunda dimensão.

Sem perda de generalidade, representando o sistema linear sobredeterminado da Equação 2.4 como:

$$\mathbf{Ax} = \mathbf{b}, \tag{2.8}$$

a solução será dada, no sentido de mínimos quadrados, pelas suas equações normais por:

$$\mathbf{x} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b}\,. \tag{2.9}$$

Os coeficientes α da matriz $\mathbf{L}$ podem ser definidos de diferentes maneiras. Paulovich *et al.* (2008) discute duas construções diferentes para α_{ij}:

$$\alpha_{ij} = \frac{1}{k_i}, \text{ (} k_i \text{ sendo a quantidade de vizinhos do ponto } \mathbf{d}_i\text{) }, \tag{2.10}$$

$$\alpha_{ij} = \left(\frac{1}{\delta(\mathbf{d}_i, \mathbf{d}_j)}\right) \left(\sum_{\mathbf{d}_k \in V_i} \frac{1}{\delta(\mathbf{d}_i, \mathbf{d}_k)}\right)^{-1} \text{ (} \delta \text{ sendo a distância euclidiana)}, \tag{2.11}$$

onde a Equação 2.11 é a utilizada nesse trabalho. Com a utilização da primeira, a matriz $\mathbf{L}$ corresponde a versão não-simétrica da matriz Laplaciana para grafos, que acaba dando a mesma importância para todos os vizinhos de um ponto, independente o quão próximo estejam - caso a vizinhança V_i seja dada por um grafo de k-NN.

Em nossos experimentos, a LSP é realizada utilizando-se os dez vizinhos mais próximos para cada *cluster* e dez pontos de controle. Para o posicionamento deles no espaço visual, utiliza-se a Force Scheme (TEJADA; MINGHIM; NONATO, 2003), visto que possui boas propriedades

de distribuição dos pontos de forma automática. Uma vez projetados os pontos de controle no espaço visual, estes são utilizados como pontos de controle para a LSP. Na Figura 16 é ilustrado um exemplo dessa etapa com o conjunto de dados IV04, que será será definido na Seção 2.4.

Figura 16 – Exemplo da projeção de documentos utilizando a técnica LSP.

Fonte: Elaborada pelo autor.

Ressalta-se que o mapeamento utilizado garante que a estrutura de vizinhança original seja preservada tanto quanto possível no espaço visual. No entanto, qualquer projeção que preserve vizinhança dos pontos pode ser adotada.

Ranking

A modificação do tamanho visual de entidades em concordância com a relevância do documento a qual elas representam é um recursos bastante útil no processo de análise e exploração de coleções de documentos.

A relevância de um documento na coleção pode ser calculada de diversas formas. No sistema aqui proposto, nos casos em que os documentos forem associados de acordo com uma dada relação, como por exemplo citações ou *hiperlinks*, essa informação define uma estrutura de grafo naturalmente, o que possibilita utilização de medidas de *centralidade em grafos* como a relevância do documento. Algumas das medidas mais comuns para centralidade em grafos são: grau, proximidade, *betweenness* e autovetor (NEWMAN, 2008).

A que mais se adéqua ao nosso problema é a por autovetor, uma vez que a sua intuição consiste em que a relevância de um nó do grafo depende da importância dos seus nós vizinhos. Aplicando isso no contexto de citações entre documentos, define que um documento é relevante se ele é citado por outros documentos relevantes, não necessariamente o que mais citações tiver. Nessa linha o algoritmo *PageRank* (LANGVILLE; MEYER, 2009) foi utilizado, que consiste no cálculo do autovetor de autovalor dominante de uma matriz estocástica.

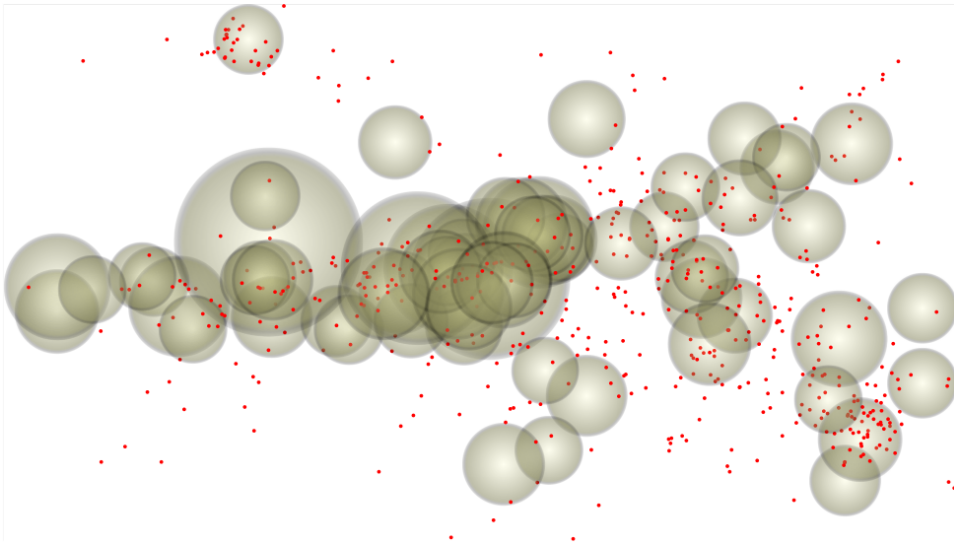
A matriz em consideração nessa etapa não é a matriz de td-idf que define os documentos, mas sim a matriz de adjacências gerada pelas referências definidas em cada artigo. Em eventuais situações onde a coleção de documentos não apresentar um citações entre os pontos do *dataset*, a relevância é calculada utilizando-se o grafo de $k-$*nearest neighbors* (k-NN) no espaço original dos documentos.

Mais formalmente, a relevância de cada documento é dada pela solução do problema de autovetor $\mathbf{Mx} = \mathbf{x}$, em que cada coluna de $\mathbf{M}$ corresponde a um documento e cada entrada $\mathbf{M}_{ij}$ assume valor não nulo quando o documento i estiver ligado ao documento j, mais precisamente, $\mathbf{M}_{ij} = 1/\text{outdeg}(i)$, onde $outdeg$ corresponde numero de arestas do i-ésimo nó.

Criação de discos e Simulação de Corpo-rígido

Cada ponto resultante da projeção multidimensional gera um corpo rígido (*rigid-body*), que no presente caso assume a forma de um disco centrado no ponto projetado e com raio definido em acordo com a relevância do documento frente a coleção. Nesse estágio, a sobreposição dos discos deve ser expressiva sendo assim necessário um alto grau de interação no acesso de grupos e indivíduos, como pode ser visto no exemplo ilustrado na Figura 17.

Figura 17 – Exemplo após a etapa de criação de discos bidimensionais para documentos, cuja raio é dado pela sua relevância na coleção.



Fonte: Elaborada pelo autor.

A simples distribuição dos discos de forma a evitar sobreposição não é efetiva, uma vez que as estruturas de vizinhança podem não ser preservadas. Por essa razão, propõem-se o uso de um esquema de força capaz de repelir uma sobreposição de discos enquanto evita a pertubação da estrutura de vizinhança inicialmente obtida.

Simular computacionalmente a dinâmica de corpos-rígidos que se intersectam, de forma

precisa, é um problema custoso (CATTO, 2005). Uma simplificação possível é proposta por Catto (2005) que utiliza-se do conceito de forças de restrição (*constraint forces*), e denominada *Box2D*. A *engine* consiste em atualizar a posição do centro de cada disco ao longo da simulação por meio de impulsos e forças agindo nos corpos. Esses impulsos são aplicados pela *engine* física responsável por evitar a interseção dos corpos, de forma que permitem lidar com velocidades ao invés de aceleração, o que torna a solução numérica mais simples, rápida e estável (CATTO, 2005). Ela permite que o usuário estabeleça vários tipos de restrições de movimento e também aplique forças a corpos de uma cena. O objetivo da *Box2D* é determinar o *estado* de cada corpo rígido de uma cena no instante $t+dt$, isto é, posição do centro de massa, orientação do sistema inercial local, velocidade linear e velocidade angular, em função do estado anterior no instante t.

No contexto da MIST, a *engine Box2D* é modificada apenas de forma a serem introduzidas forças de atração, como *força aplicada* $\mathbf{f}^a$ nos discos, já que não consideramos gravidade, como forma de se evitar que discos vizinhos se distanciem durante a simulação. A força de atração $\mathbf{f}_i^a$ associada a um disco i é calculada como:

$$\mathbf{f}_i^a = \sum_{j \in N_i} m_i m_j d_{ij} \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \tag{2.12}$$

em que $m_i = \pi r_i^2$ é a massa e r_i é o raio do disco, $d_{ij} = \max\left\{0, \|\mathbf{x}_i - \mathbf{x}_j\|_2 - \left(r_i + r_j\right)\right\}$, e N_i corresponde à vizinhança no espaço visual obtida pelos k-NN do disco. Em nossas simulações, o valor de k é escolhido ser $\sim 2\%$ do número total de documentos. As forças de restrição nas posições e na velocidade, que determinam os impulsos que movem os corpos, são calculadas automaticamente pela *engine*.

Para ter uma intuição mais especificamente de como ela efetua isso, há vários tipos de restrições (*constraints*) possíveis como: posição, contato, velocidade, impulso. Diz-se que há uma *restrição de posição* entre um ponto qualquer do plano $\mathbf{p}$ e um disco de centro $\mathbf{x}_i$, raio r_i quando:

$$C(\mathbf{p}, \mathbf{x}_i) = \|\mathbf{p} - \mathbf{x}_i\| - r_i = 0, \tag{2.13}$$

o que implica dizer que as soluções são os pontos $\mathbf{p}$ do plano que estejam restritos à borda do disco. Note que pode ser observada como uma representação implícita do disco. ,

Através dessa restrição, utilizando a regra da cadeia, e tomando $\mathbf{p} = \mathbf{x}_j - \mathbf{x}_i$, para simplificar as contas, sendo disco j que intersecta o disco i, deriva-se a *restrição de velocidade* ficando definida como:

$$\dot{C} = \frac{dC}{dt}(\mathbf{p}) = \left\langle \frac{\mathbf{p}}{\|\mathbf{p}\|}, \mathbf{v}_j \right\rangle = \left(\frac{\mathbf{p}}{\|\mathbf{p}\|}\right)^\top \mathbf{v}_j = 0, \tag{2.14}$$

considerando o disco i como referência (origem). Aqui utiliza-se que os discos possuem velocidade angular nula, ou seja, suas rotações não modificariam a intersecção dos discos. A restrição implica que o disco j tem velocidade linear nula na direção normal de contato, que seria dada por $\mathbf{p}$[1]. Ao considerar o caso mais geral para n corpos Catto (2005) mostra que a Equação 2.14

[1] Essa construção pode acabar levando ao efeito de *stacking*, citado posteriormente.

passa a ser:

$$\dot{C} = \mathbf{J}\mathbf{V} = 0, \tag{2.15}$$

onde $\mathbf{J}$ é a matriz jacobiana e $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)^\top$ um vetor coluna formado pelas $2n$ velocidades lineares dos discos considerados. Essa restrição implica que a velocidade admissível na solução tem que ser ortogonal às linhas da matriz $\mathbf{J}$, com isso a *engine* então aplica impulsos em cada uma dessas direções. Assim,

$$\mathbf{f}_c = \mathbf{J}^\top \lambda, \tag{2.16}$$

onde λ é um vetor com as magnitudes a serem determinadas para cada disco considerado. Com isso, para o discos em posição inicial ($t = 0$) e de centros $\{\mathbf{x}_i^0\}$, a *engine* efetua basicamente as seguintes etapas:

(1) Definição de forças

a) As forças agindo nos discos são separadas em *forças aplicadas* $\mathbf{f}^a$ e *forças das restrições* $\mathbf{f}^c$, sendo agrupadas nas matrizes $\mathbf{F}^a$ e $\mathbf{F}^c$ respectivamente, logo:

$$\mathbf{M}\dot{C} = \mathbf{F}^a + \mathbf{F}^c, \tag{2.17}$$

$\mathbf{M}$ a matriz de massas dos discos, com $m_{ii} = \pi r_i^2$;

(2) Correção da velocidade para as forças aplicadas

a) Integração das *forças aplicadas* utilizando o esquema de Euler explícito

$$\hat{\mathbf{V}}^{t+1} = \mathbf{V}^t + h\mathbf{M}^{-1}\mathbf{F}^a \ , \tag{2.18}$$

(3) Correção da velocidade para as forças de restrição,

a) Cálculo do vetor λ

$$\lambda = -(\mathbf{J}\hat{\mathbf{V}}^{t+1})^{-1}(\mathbf{J}\mathbf{M}^{-1}\mathbf{J}^\top) \ , \tag{2.19}$$

b) Definição dos impulsos para cada restrição c (*constraint impulse*)

$$\mathbf{P}_c = h\mathbf{F}^c = h\mathbf{J}^\top \lambda, \tag{2.20}$$

c) Integração das *forças de restrições*, utilizando os momentos $\mathbf{P}_c$

$$\mathbf{V}^{t+1} = \hat{\mathbf{V}}^{t+1} + \mathbf{M}^{-1}\mathbf{P}_c \ . \tag{2.21}$$

(4) Correção da posição do centro

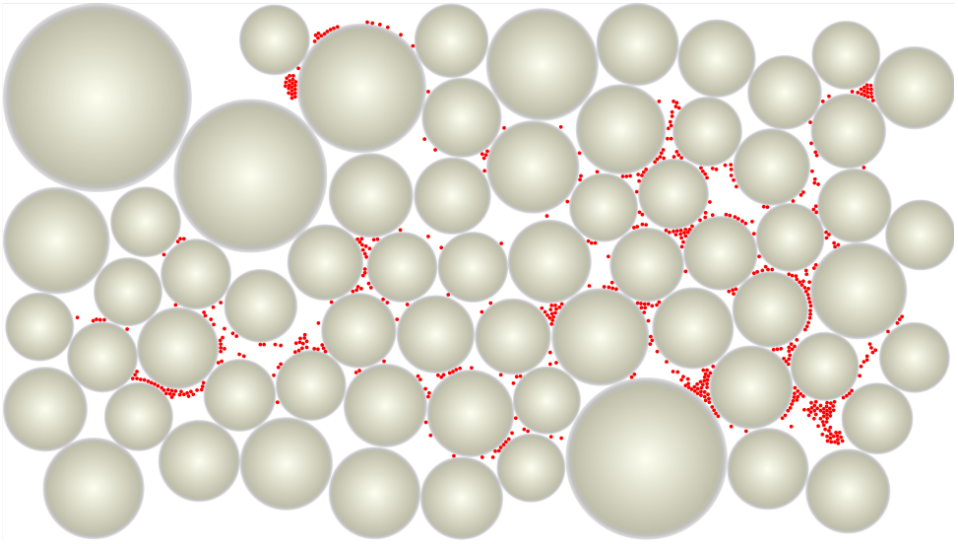
a) Integração das novas velocidades utilizando o esquema de Euler semi-implícito para atualizar as posições dos discos, já que utiliza o termo de velocidade no tempo $t+1$, na forma:

$$\mathbf{X}^{t+1} = \mathbf{X}^t + h\mathbf{V}^{t+1} \ . \tag{2.22}$$

A escolha por essa *engine* em particular deve-se a sua estabilidade quando comparada a sistemas massa-mola. Ela tem sido utilizada em jogos virtuais devido sua performance e estabilidade em detrimento de precisão numérica. Além disso, ela funciona bem tanto quando os corpos são discos como polígonos convexos mais gerais.

As fronteiras da janela principal podem também ser consideradas como corpos rígidos, de forma a forçar que os discos mantenham-se confinados na região visual. Esse recurso pode ou não ser habilitado no presente sistema. Um exemplo do resultado desse processo pode ser observado na Figura 18.

Figura 18 – Exemplo após a distribuição dos discos pelo plano, removendo possíveis intersecções.



Fonte: Elaborada pelo autor.

Ao fim dessa etapa é possível identificar de forma mais aparente duas coisas: a distribuição de relevância entre os documentos na coleção; documentos cuja relevância associada é menor do que os exibidos como disco, o que possibilita uma noção de densidade na coleção e exploração multi-escala. Detalhes sobre esse processo são descritos no final da Seção 2.3.

Agrupamento de discos

Após os discos estarem propriamente dispostos de acordo com a simulação, para que seja criado um sumário do conteúdo de cada documento que reflita a sua nova disposição no espaço visual é efetuada uma etapa de agrupamento (*clustering*), de forma que cada agrupamento será utilizado como base para criação da nuvem de palavras na etapa seguinte.

Formas automáticas de agrupamento poderiam ser consideradas nessa etapa, por ex., levando em consideração a distribuição calculada de relevância dos discos, agrupamento por densidade. Porém, para possibilitar flexibilidade ao usuário, utiliza-se o *k-means*++ (ARTHUR; VASSILVITSKII, 2007) ao grupo de discos, em que o número k de *clusters* é um parâmetro livre,

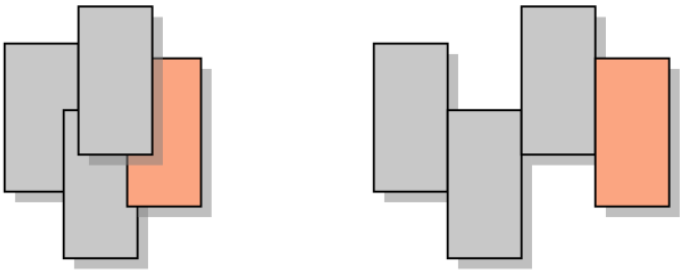
informado pelo usuário. Com isso, ele pode controlar a granularidade da sumarização efetuada após distribuição dos discos, aumentando a flexibilidade da análise efetuada.

É importante salientar que o agrupamento utiliza-se da disposição dos discos no espaço visual após a remoção de sobreposição, uma vez que a relação de vizinhanças pode ser alterada durante esse processo. Essa utilização justifica-se para que posterior construção da sumarização utilizando nuvem de palavras, faça sentido, ou seja, palavras vizinhas correspondam a documentos vizinhos, na metáfora proposta.

Nuvem de palavras multi-semente

Uma estratégia inicial de construção das nuvens de palavras poderia utilizar as caixas envolventes (*bounding box*) de cada palavra e efetuar seu posicionamento utilizando a engine *Box2D*. Entretanto, Strobelt *et al.* (2012) mostra que a *engine* possui um efeito de empilhamento (*stacking*) quando o corpo rígido a ser posicionado possui tamanhos diferentes em cada dimensão, isto é, retângulos, ilustrado na Figura 19.

Figura 19 – Efeito de empilhamento produzido pela *Box2D* para corpos rígidos com diferentes tamanhos em cada dimensão, isto é, retângulos.

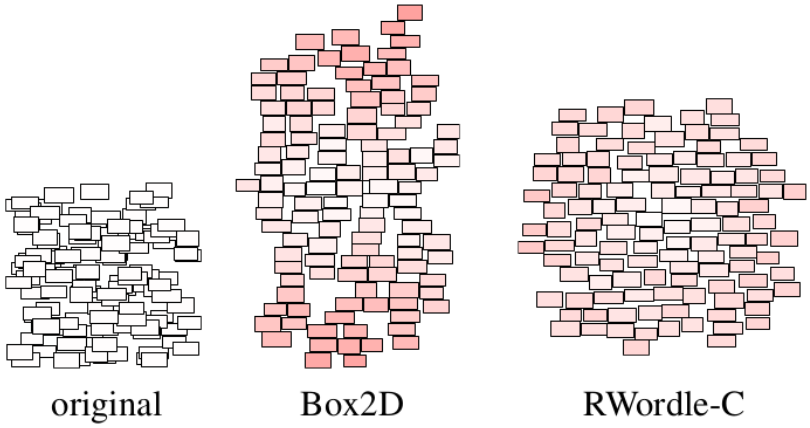


Fonte: Adaptada de Strobelt *et al.* (2012).

Cabe notar aqui que esse efeito não foi verificado quando os tamanhos são iguais para as dimensões, que ocorre no caso descrito dos discos. Strobelt *et al.* (2012) compara sua distribuição com *Box2D*, ilustrando como apresenta ganhos significativos na distribuição dos corpos rígidos, ilustrado na Figura 20.

Para contornar esse problema, foi efetuada uma construção inspirada no esquema espiral utilizado por Strobelt *et al.* (2012), com sua construção descrita a seguir. O centro da caixa envolvente (*bounding box*) de cada *cluster* c_i é utilizado para iniciar o processo de construção de sua nuvem de palavras correspondente. O conjunto de palavras W_i associado ao *cluster* c_i é classificado de forma decrescente em relação a relevância. A palavra-chave mais relevante é disposta no centro da caixa envolvente b_i do *cluster* c_i. O tamanho da fonte é definido de acordo com o comprimento da caixa envolvente b_i. Partindo-se do máximo tamanho de fonte permitido da palavra-chave mais significante, a fonte de uma dada palavra-chave decresce até

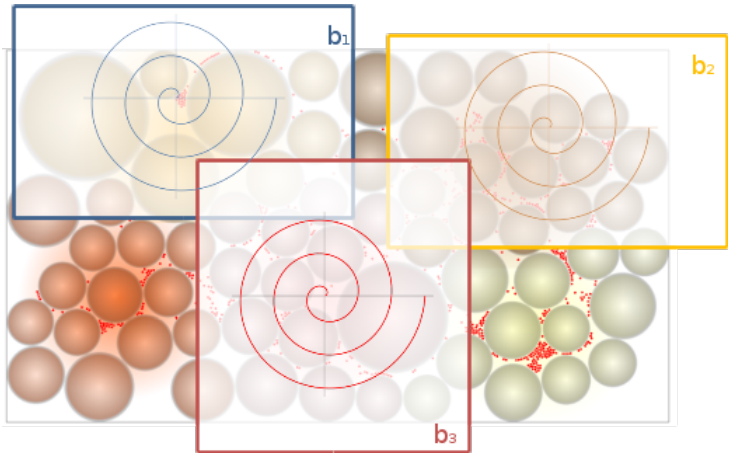
Figura 20 – Comparação entre distribuição de corpos-rígidos em forma de retângulo pelas técnicas *Box2D* e uma das abordagens do RWordle (STROBELT *et al.*, 2012). Em vermelho é quantificado o desvio da posição inicial de cada corpo-rígido.



Fonte: Adaptada de Strobelt *et al.* (2012).

atingir o valor de 70% do comprimento de b_i. O tamanho de fonte das demais palavras-chave é linearmente interpolado entre o maior e menor tamanho de fonte. As palavras-chave são então horizontalmente dispostas no interior de b_i seguindo o procedimento espiral proposto por RWordle (STROBELT *et al.*, 2012). O menor tamanho de fonte é pré-fixado. O procedimento é finalizado quando não houver espaço disponível em b_i para inserir uma palavra-chave. O posicionamento espiral é efetuado para cada palavra-chave, dentro de cada cluster b_i, como ilustrado na Figura 21.

Figura 21 – Posicionamento de palavras-chave em espiral dentro de cada *cluster*, e suas respectivas caixas envolventes (*bounding-boxes*).



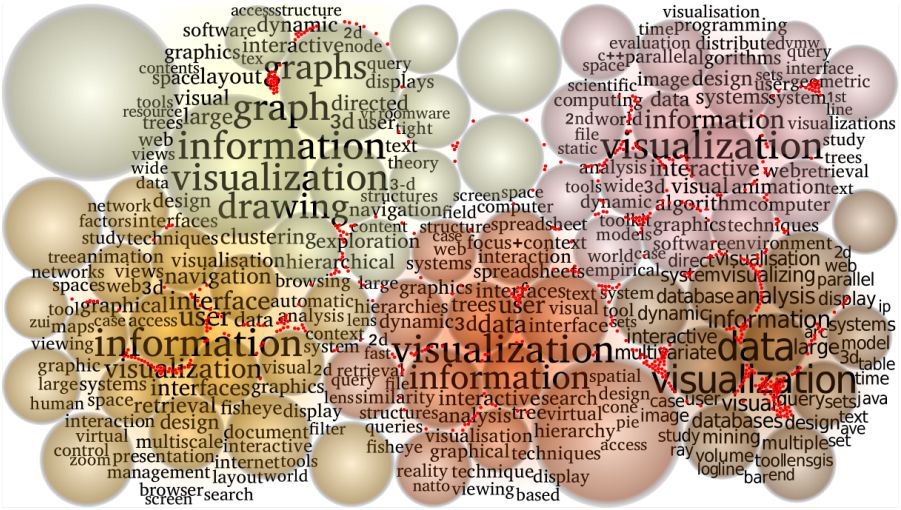
Fonte: Elaborada pelo autor.

Uma vez que *bounding boxes* de *clusters* distintos podem se sobrepor, a caixa que contém uma palavra-chave a ser inserida em b_i pode interceptar a caixa de palavras-chaves em outros *clusters*. De forma a acelerar o processo de verificação de interseções entre palavras-chave em *clusters* distintos, utiliza-se uma árvore dinâmica para armazenar os retângulos contendo

cada palavra-chave. A estrutura de dados adotada é uma variação de uma *bounding volume tree* dinâmica de Presson utilizada no motor de física Bullet (COUMANS, 2012). Ao fim do processo, será obtida uma distribuição como a ilustrada na Figura 22.

Usuários podem também modificar a fonte e a cor utilizada por cada nuvem de palavras. Palavras-chaves associadas a cada documento, obtidas na etapa de pré-processamento, são filtradas de acordo com sua relevância. A relevância de uma palavra-chave é dada em função do número de ocorrências de cada palavra-chave presente nos documentos que compõem o *cluster*. Vale salientar que a tarefa de *stemming* não é efetuada para a geração das nuvens de palavras. Além disso, para sua definição são utilizados os títulos dos artigos como elemento textual, sem a etapa de *stemming*.

Figura 22 – Resultado obtido após estratégia de posicionamento das palavras-chave em espiral, por *cluster*.



Fonte: Elaborada pelo autor.

Como ilustrado na Seção 2.4, o sistema também permite que a informação de relevância de palavras-chaves seja um argumento de entrada, o que possibilita o uso de técnicas mais sofisticadas e que preservem semântica (e.g., Paulovich *et al.* (2012)).

Aspectos computacionais

O tamanho dos discos que representam cada documento e a unidades de medida utilizadas na simulação de corpo rígido são derivados a partir da definição da janela de exibição. O documento mais relevante fica representado pelo disco de maior raio $r_{\max}$ enquanto o disco que representa o documento menos relevante possui raio $r_{\min}$. Os raios $r_{\max}$ e $r_{\min}$ são definidos de forma que as áreas do maior e menor disco correspondam a 5% e 0.5% da área total, respectivamente. O raio de todos os demais discos são linearmente interpolados entre $r_{\max}$ e $r_{\min}$ respeitando-se o respectivo valor de relevância. O número de discos a ser visualizado também é definido de acordo com a área da janela de exibição, ou seja, a área dos discos é acumulada

em ordem descendente até que a mesma atinja 75% da área da janela. Os primeiros t discos restantes, escolhidos de acordo com as respectivas relevâncias, são exibidos em uma pequena área. Na simulação física, instâncias menos relevantes não são visíveis e acabam por se moverem como uma única instância junto ao disco visível mais próximo. Esse valor de t pode ser ajustado mas, por padrão, assume como valor duas vezes o número de discos visíveis.

Na situação de um usuário selecionar uma região retangular na janela de exibição, para uma exploração mais detalhada (*navigation zoom*), a região selecionada é mapeada para uma nova janela de exibição. Todo o processo descrito anteriormente é executado nessa nova configuração e é restrita ao subconjunto de instâncias contidas na região selecionada pelo usuário, como ilustrado na Figura 9.

Outro aspecto a ser discutido refere-se a como aumentar a velocidade de execução da simulação física por meio de mudanças na unidade de comprimento. A ferramenta Box2D é otimizada a corpos com tamanhos variando de uma unidade de medida (CATTO, 2005). Assim, as coordenadas dos centros e raios de cada disco são escalados de forma a pertencerem a esse intervalo com essa magnitude dessa variação citada.

2.4 Resultados e comparações

De forma a explicitar a performance da estratégia MIST, realiza-se testes e um comparativo utilizando três bases de dados. Uma base de dados inclui todos os trabalhos (616 documentos) publicados na IEEE Information Visualization Conference (IV) de 1995 até 2002 (FEKETE; GRINSTEIN; PLAISANT, 2004), juntamente com metadados como título, palavras-chaves, resumo e referências. Utiliza-se também a base de dados High Energy Physics[2], a qual contém informações de resumos e citações de artigos na área de física teórica de altas-energias. Especificamente, divide-se a base de dados original em dois subconjuntos com 2000 (HEP2) e 3000 (HEP3) instâncias obtidas randomicamente.

Uma avaliação da efetividade da abordagem proposta no que se refere as métricas de organização do *layout*, preservação de vizinhança e compactabilidade, é realizada a partir de um conjunto de comparações, junto a quatro métodos baseados em força (*force-directed*), tal como segue:

1. O método de Fruchterman-Reingold (FR) (FRUCHTERMAN; REINGOLD, 1991) emprega um modelo massa-mola de forma a colocar nós de um grafo no espaço visual utilizando um número reduzido de parâmetros;

2. O método de Yifan Hu (YH) (HU, 2005) faz uso de uma abordagem que aplica um esquema multinível de forças;

[2] kdl.cs.umass.edu/data/hepth/hepth-info.html

3. O método *Attractive and Repulsive Forces* (ARF) (GEIPEL, 2007) amplia a metodologia *spring-based* no sentido de reduzir vazios no *layout*.

   Essas três técnicas permitem a representação dos nós como discos. No entanto, essas técnicas apenas evitam a interseção dos discos para o caso em que o raio é constante, ou seja, para o caso em que todos os discos apresentarem o mesmo tamanho. A interseção entre discos com raios distintos não pode ser controlada por meio do ajuste de parâmetros;

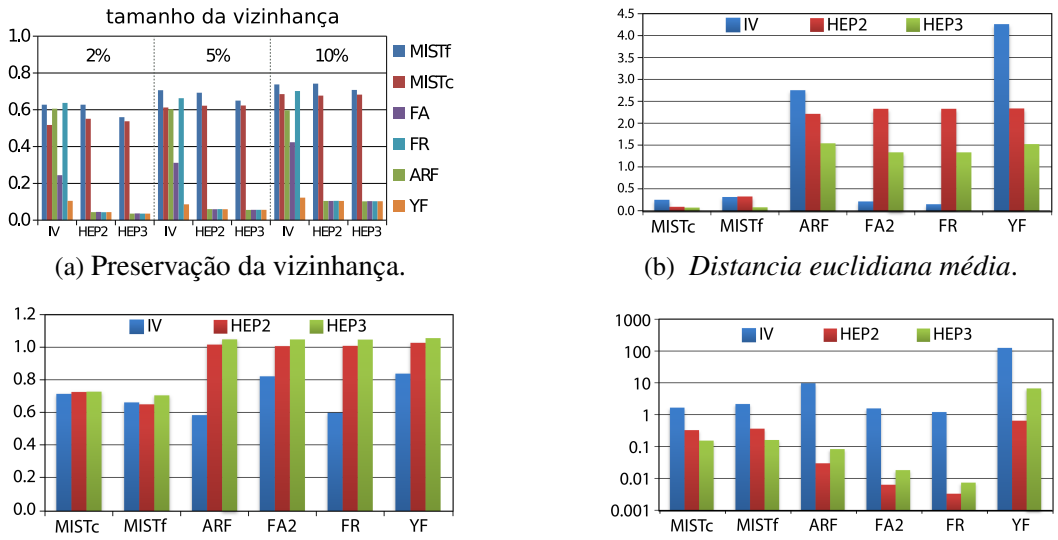
4. Em contraste às técnicas FR, YH e ARF, a técnica ForceAtlas2 (FA2) (GIBSON; FAITH; VICKERS, 2012) é capaz de gerenciar a intersecção de discos de raios variados. A técnica faz uso de forças lineares de atração e repulsão de forma a dispor discos no espaço visual enquanto evita sobreposição.

Quatro métricas distintas foram utilizadas para se comparar a técnica MIST com as técnicas anteriormente citadas: preservação de vizinhança (k-NN), preservação média de distância euclidiana, dissimilaridade do *layout* e aumento de tamanho geral, sendo as três últimas propostas por Strobelt *et al.* (2012).

A métrica de *preservação de vizinhança* calcula a porcentagem média da quantidade de vizinhos do espaço original que se manteve na vizinhança no espaço visual, após todo o processo. Os resultados obtidos podem ser vistos na Figura 23a.

A métrica de *preservação de distância euclidiana* é dada por $E = \frac{1}{n}\sum_i \|\mathbf{x}_i^o - \mathbf{x}_i\|_2$, onde $\mathbf{x}_i^o$ e $\mathbf{x}_i$ são a posição inicial e final do centro dos discos, n é o número de discos. Esse valor mede o deslocamento do discos durante a simulação sabendo-se que um deslocamento pequeno

Figura 23 – Análise quantitativa de preservação de vizinhança e compacidade do *layout* das técnicas MIST, FR, ARF, FA2 e YH.



(a) Preservação da vizinhança.

(b) *Distancia euclidiana média.*

(c) *Dissimilaridade de Layout.*

(d) *Aumento de tamanho.*

Fonte: Elaborada pelo autor.

implica em uma preservação da configuração inicial. Os resultados obtidos podem ser vistos na Figura 23b.

A métrica de *dissimilaridade do layout* quantifica o grau com que a estrutura de vizinhança é afetada por meio da simulação. A ideia baseia-se em se medir o quanto o comprimento das arestas do grafo de k-NN inicial é modificado depois da simulação. Em termos matemáticos, sendo l_k^o e l_k os comprimentos das arestas antes e depois da simulação, a métrica é dada por

$$\sigma = \frac{1}{\bar{r}}\sqrt{\frac{1}{m}\left(\sum_{k=1}^{m}(r_k - \bar{r})^2\right)}, \quad \bar{r} = \frac{1}{m}\sum_{k=1}^{m} r_k \ , \tag{2.23}$$

onde $r_k = l_k/l_k^o$ e m é o número de arestas no grafo k-NN. Valores pequenos da métrica correspondem a uma boa preservação do *layout*. Os resultados obtidos podem ser vistos na Figura 23c.

Finalmente, dados o fecho convexo C^o e C dos *layouts* original e modificado, a *métrica de aumento no tamanho* é calculada como $S = \text{area}(C)/\text{area}(C^o)$ e ela determina a mudança relativa tanto em tamanho como compacidade do *layout* modificado. Valores próximos a um são desejáveis e implicam que a técnica adotada não expande nem compacta a área designada originalmente. Os resultados obtidos podem ser vistos na Figura 23d.

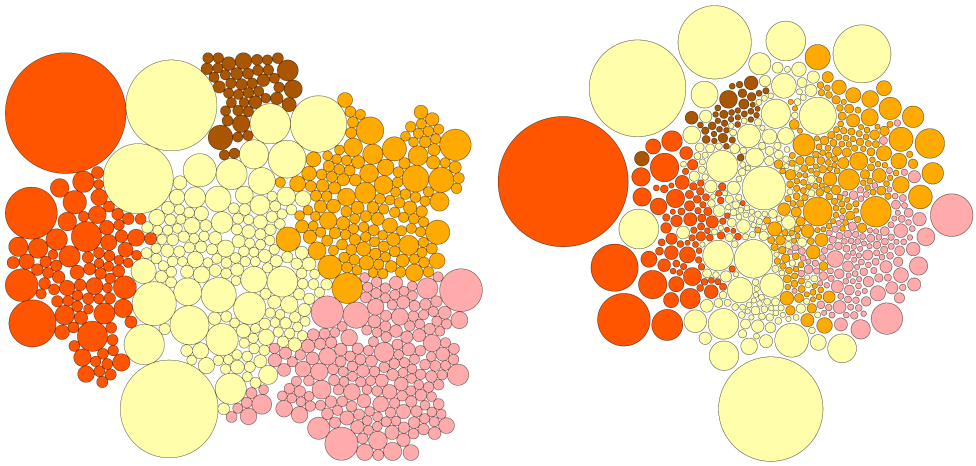
A Figura 23 apresenta resultados obtidos pelas métricas escolhidas quando aplicadas nos *layouts* produzidos pela MIST e pelos métodos utilizados como comparativo, quando utilizados na visualização das bases de dados IV, HEP2 e HEP3. Assume-se uma mesma configuração inicial, obtida pela projeção multidimensional.

A relevância da informação não foi levada em consideração na análise, com o objetivo de se manter iguais os raios de todos os discos. A hipótese em questão é condição necessária para que os métodos FR, YH e ARF tratem a sobreposição dos discos adequadamente. Uma vez que MIST permite o confinamento do *layout* em uma caixa, testes foram realizados utilizando as duas versões do algoritmo. *Constrained* MIST (MISTc) é definido como o *layout* com confinamento e MISTf representa casos onde o confinamento é desabilitado. Observa-se que MIST apresenta uma performance significativamente melhor que os outros métodos para todas as métricas quando sujeitos as grandes bases de dados HEP2 e HEP3. Resultados razoáveis são também obtidos com a base IV. Os resultados quantitativos apresentados na Figura 23 atestam que MIST produz resultados mais compactos além de bem preservar a estrutura de vizinhança inicial provida pela projeção multidimensional (Multidimensional Projection - MP).

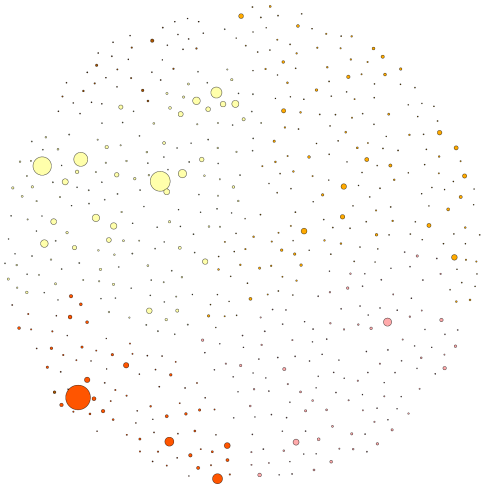
A Figura 24 mostra que o *layout* resultante pelos métodos da MIST, FA2 e FR ao se visualizar a base de dados IV levando-se em consideração a informação de relevância.

O uso de cores indica de que forma vizinhanças são preservadas após a simulação. Mais especificamente, aplica-se primeiramente *k-means* no sentido de se clusterizar, no espaço visual, as saídas da projeção. Na sequência, executa-se a simulação com o objetivo de se verificar o quão bem os *clusters* (estruturas de vizinhança) são preservadas. Resultados claramente demostram

Figura 24 – *Layouts* gerados pelas técnicas MIST, FA2 e FR. MIST apresenta uma melhor preservação dos *clusters*.



(a) MIST. (b) FA2.



(c) FR.

Fonte: Elaborada pelo autor.

que a MIST melhor preserva os *clusters* e os mantém um *layout* compacto. Ao contrário do que o ocorre com a MIST, no qual demanda apenas o número de *clusters* como parâmetro, o usuário necessita definir um conjunto de vários parâmetros para que o método FR consiga tratar a sobreposição. Como pode ser observado em 24c, a busca por um conjunto de parâmetros que forneça um resultado em *layout* compacto e sem sobreposição não é tarefa fácil.

A Figura 25 compara as nuvens de palavras geradas pela MIST e as geradas pelo mecanismo de *wordification* proposto por Paulovich *et al.* (2012).

Enquanto o método de Paulovich utiliza espaços em branco para separar grupos de nuvens de palavras, MIST emprega cores de fundo que permitem, em uma representação mais compacta, a identificação grupos de distintas as nuvens de palavras. Para essa comparação, empregou-se o mesmo número de *clusters* e palavras-chaves para ambos os métodos. A relevância de cada palavra foi também obtida por (PAULOVICH *et al.*, 2012).

Figura 25 – Comparativo de nuvens de palavras geradas pelo (a) método de Wordification (PAULOVICH *et al.*, 2012) e (b) MIST, utilizando o mesmo número (5) de *clusters* e o mesmo conjunto de palavras-chaves.

(a) Wordification.

(b) MIST.

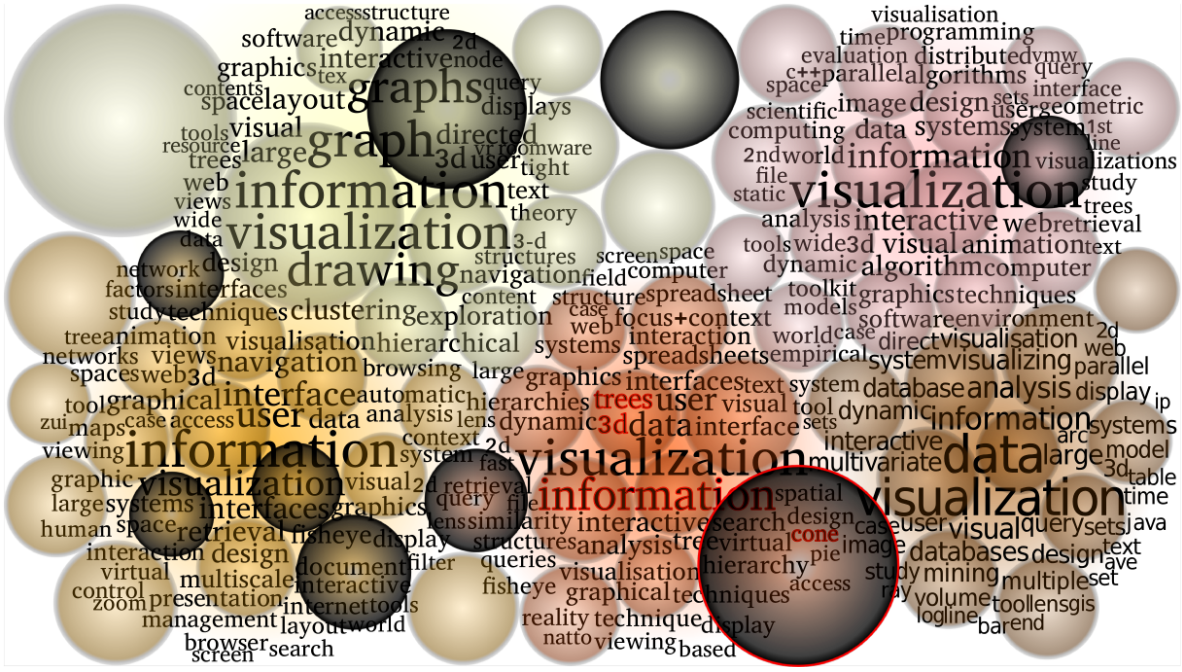
Fonte: Elaborada pelo autor.

De forma geral, duas importantes funcionalidades da MIST são destacadas: a visualização da relação de citação entre os documentos e uma navegação exploratória.

A primeira, ilustrada na Figura 26, a MIST enfatiza a visualização de *links* entre documentos. Em um exemplo específico da coleção de documentos científicos, quando o usuário seleciona um documento (*red ring*), todos os documentos que o citam serão também destacados, bem como as palavras-chave relativas são destacadas em vermelho. A funcionalidade é significativamente útil no processo de análise de referências em artigos científicos e redes de citações assim como em páginas Web. É possível identificar rapidamente que o artigo selecionado refere-se a visualização de *cone trees 3d*, por serem as palavras destacadas em vermelho, sendo também as do artigo que estão na nuvem de palavras correspondente.

No experimento ilustrado na Figura 27, pode ser visto um comportamento interessante.

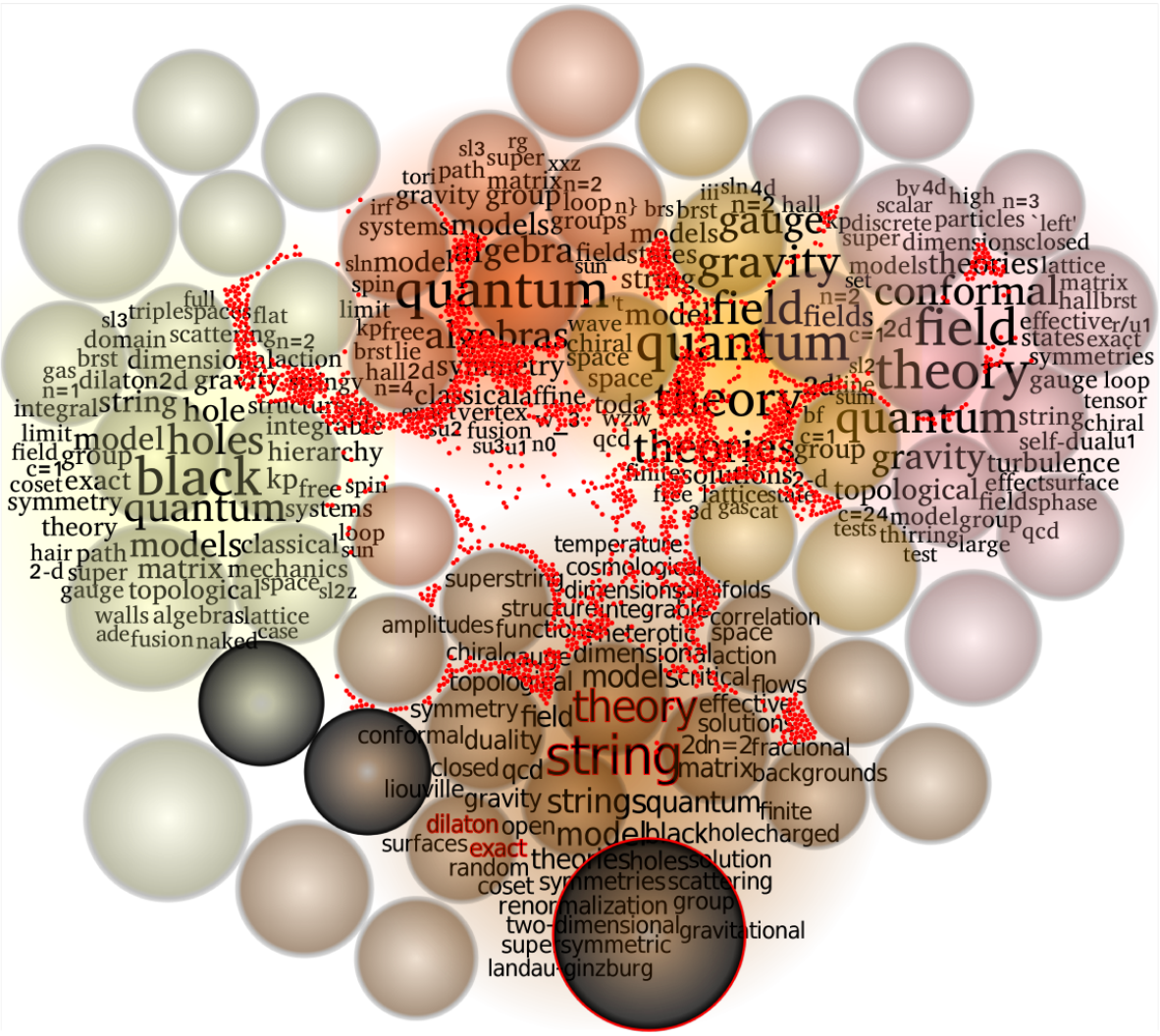
Figura 26 – Destacando a informação de citação entre artigos.



Fonte: Elaborada pelo autor.

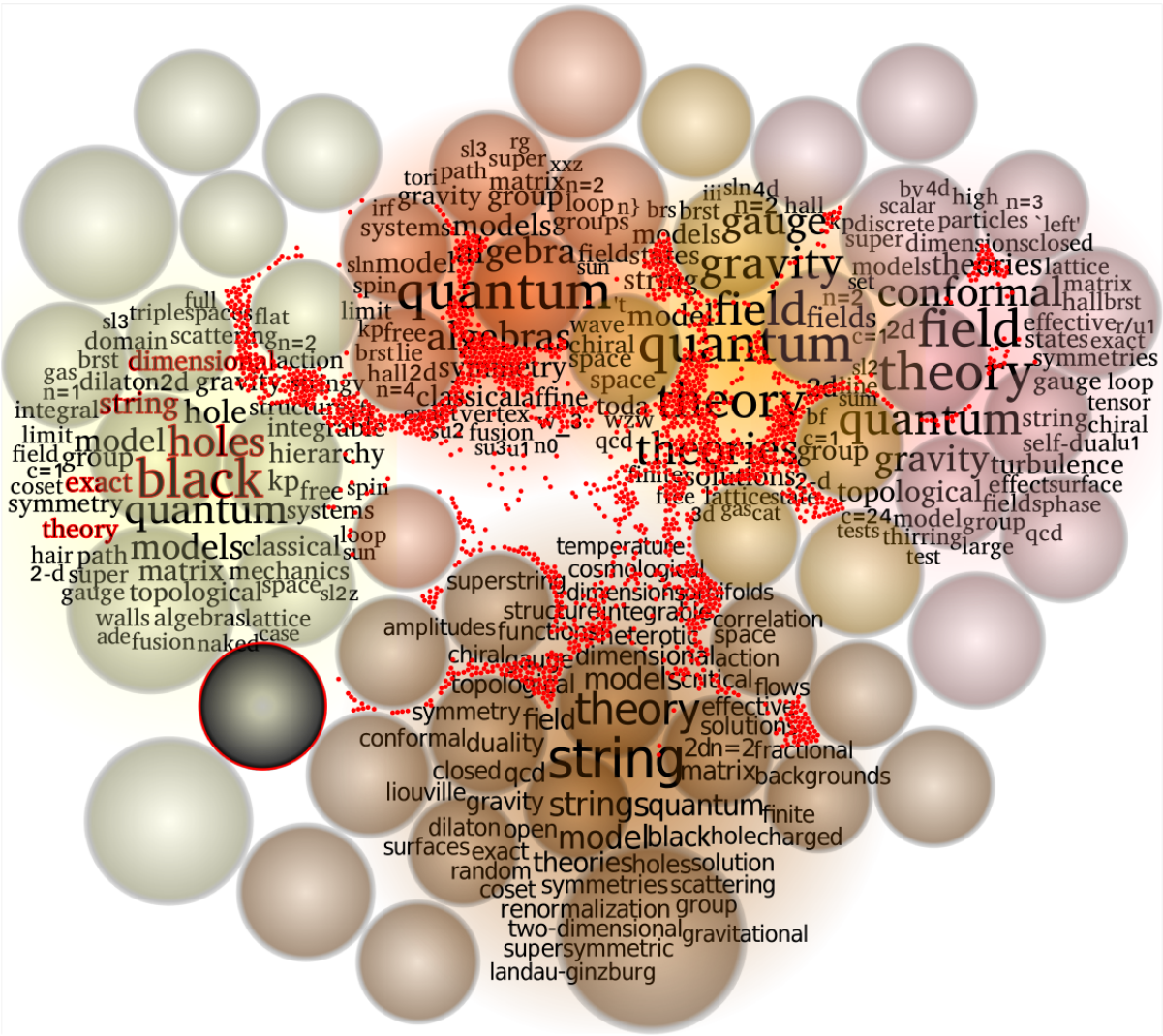
Utilizando o conjunto de dados HEP2000, ao selecionar um documento na área de teoria das cordas (*string theory*), intitulado "*Conformally exact metric and dilation in string theory on curved spacetime*", observa-se imediatamente citações que possui de documento de uma outra área, buracos negros (*black holes*), na interface entre os clusters, intitulado "*Exact Three Dimensional Black Holes in String Theory*", cuja seleção é ilustrada na Figura 28, destacando as suas palavras correspondentes na sumarização.

Figura 27 – Experimento com o conjunto de dados de HEP2000, com um documento do *cluster* de teoria das cordas selecionado.



Fonte: Elaborada pelo autor.

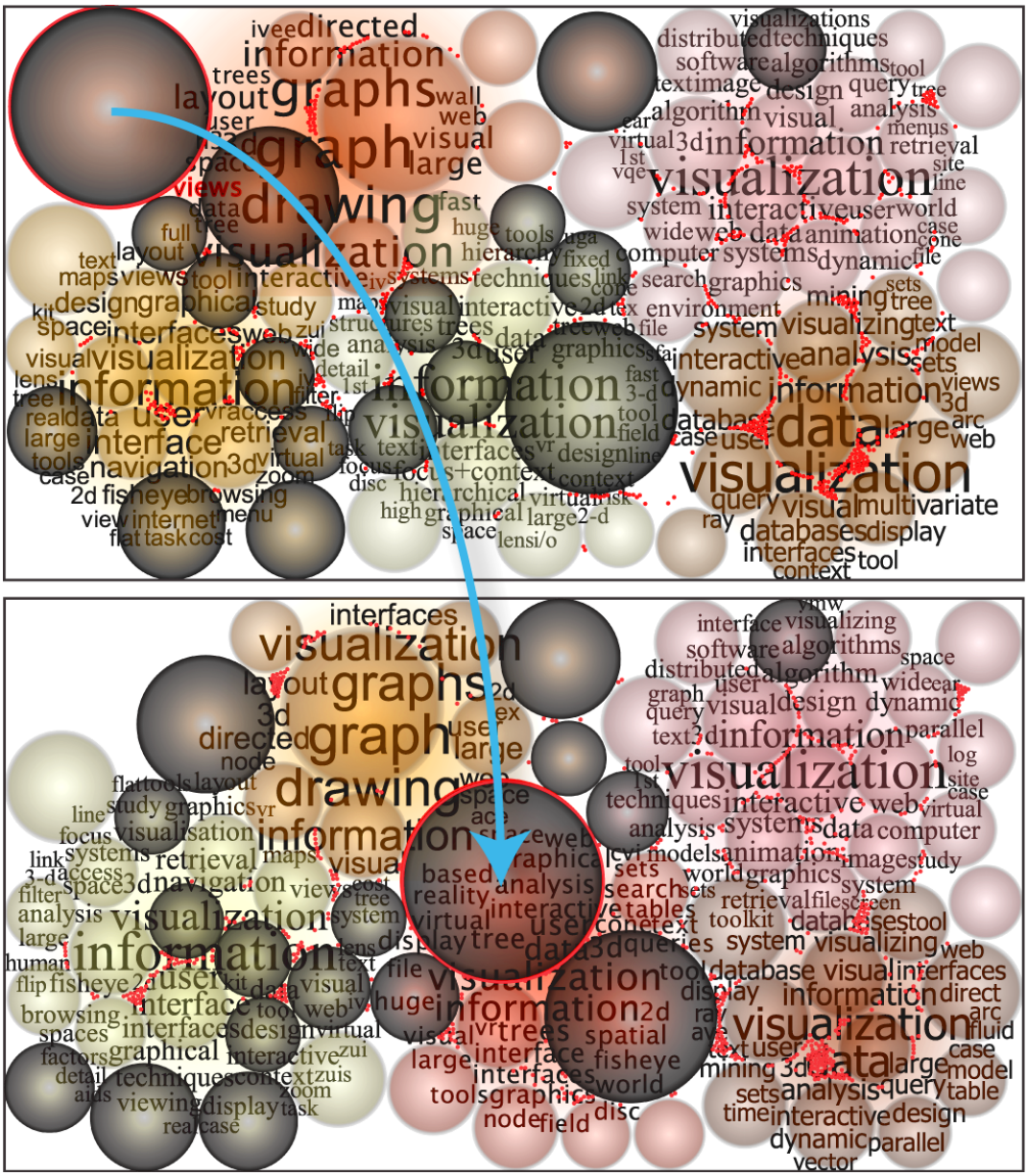
Figura 28 – Experimento com o conjunto de dados de HEP2000, com um documento do *cluster* de buracos negros selecionado.



Fonte: Elaborada pelo autor.

A Figura 29 mostra que a possibilidade de utilizar o conhecimento prévio do usuário nas áreas da coleção de documentos, através de um melhoramento na definição da similaridade entre os documentos. Possibilita também que ele efetue eventuais correções no processo de movimentação dos discos e de agrupamento, de acordo com seus critérios. Para isso, ele pode arrastar documentos de forma a modificar o *layout* e gerar disposições alternativas. Isso possibilita a modificação pelo usuário da relação de similaridade calculada, com o objetivo de melhorar a definição de similaridade dos documentos, calculada na etapa de pré-processamento. Uma vez que discos terão suas posições redefinidas após essa interação, a *engine* deve ser executada novamente e os *clusters*,que servem de base para as nuvens de palavras, serão recriados. Sumarizando a nova distribuição induzida por esse processo.

Figura 29 – Interação com o usuário através do reposicionamento de documentos.

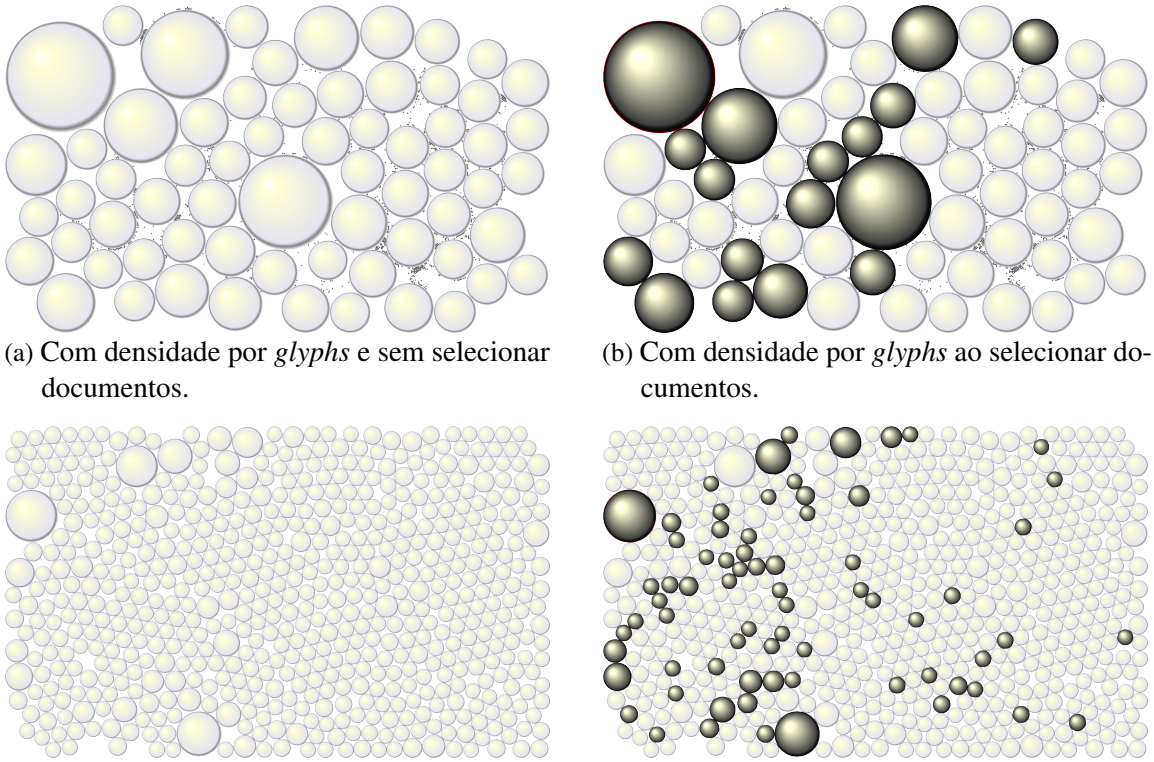


Fonte: Elaborada pelo autor.

A segunda funcionalidade é o *zoom* exploratório, que já foi ilustrado na Figura 9. Os pequenos *glyphs* que se movem com o disco ajudam a transmitir a ideia de densidade de cada *cluster*. A informação de densidade é útil ao indicar quais regiões da visualização escondem informação e devem ser exploradas com mais detalhes. A visualização desse conjunto de informações, escondidas devido a baixa relevância, é acessível ao usuário se o mesmo desenhar um retângulo definindo a nova região a ser explorada. Para esses casos, MIST reinicia o processo de visualização a partir dos dados contidos na região definida pelo usuário. Observa-se na Figura 9 que a navegação pode ser realizada recursivamente, possibilitando uma investigação detalhada da coleção de documentos. É ainda importante destacar que o contexto não é perdido durante a navegação uma vez que as janelas de exploração previamente definidas são mantidas pela visualização até que a respectiva janela seja fechada.

Essa abordagem multiescala com a estratégia de utilização dos *glyphs* afeta consideravelmente na percepção de relevância relativa entre os discos. Observe na Figura 30, onde é feita essa comparação da estratégia com e sem o uso dos *glyphs*. Ao utilizar essa estratégia observa-se uma distinção mais fácil entre as relevâncias dos documentos mais relevantes da coleção.

Figura 30 – Comparação do uso de *glyphs* contra visualização de toda a coleção.



(a) Com densidade por *glyphs* e sem selecionar documentos.

(b) Com densidade por *glyphs* ao selecionar documentos.

(c) Sem densidade por *glyphs* e sem selecionar documentos.

(d) Sem densidade por *glyphs* ao selecionar documentos.

Fonte: Elaborada pelo autor.

2.5 Limitações da técnica

Nos experimentos aqui conduzidos é possível observar que FR e FA2 não convergem para certas configurações iniciais, obtidas pela etapa de projeção, isto é, não conseguem remover a sobreposição totalmente entre os discos. Adicionalmente, YH tende a produzir *layouts* alongados quando parâmetros são modificados para evitar que discos de tamanhos variados se sobreponham. Nenhuma dessas limitações foram observadas nos experimentos realizados com a MIST.

Um comparativo justo entre as técnicas em termos de tempo de execução não é uma tarefa simples, uma vez que FR, ARF, FA2 e YH são implementadas em Java e o sistema MIST é completamente codificado em C++. No entanto, de forma a fornecer uma visão geral em termos de tempo de processamento, a MIST fornece visualizações em taxas interativas, já que o método processa apenas dados que estarão presentes na janela de visualização. Os outros métodos, no entanto, demandam dezenas de segundos para processar as bases de dados HEP2 e HEP3, limitando interatividade.

O mecanismo responsável por tornar MIST escalável consiste em deslocar documentos não exibidos juntamente com seus vizinhos mais próximos, reduzindo a quantidade de avaliações da *Box2D*. Porém, a depender do número de discos visíveis, pode acabar modificando a estrutura de vizinhança. Embora as forças de atração tendam a trazer instâncias vizinhas próximas umas as outras, alguns discos visíveis podem ser colocados entre dois vizinhos ocultos. Nesse caso, a força de atração ligada aos discos vizinhos pode não ser efetiva e a estrutura de vizinhança pode vir a ser modificada. Alternativas visando sobrepor essa limitação estão sendo investigadas.

A dificuldade na obtenção de bases de dados com citação entre documentos pode constituir uma limitação na aplicabilidade da MIST. Formas automáticas de criação de tais bases através de mineração em bases de artigos indexados ou em sites de busca de artigos em geral não estão disponíveis para uso público.

Uma forma de contornar isso é através da criação manual de tais bases, o que pode ser um processo bastante tedioso e com uma maior chance de introdução de erros. Além disso, o tamanho da base criada pode tornar a informação de relevância de cada documento menos expressiva, uma vez que na base construída pode não conter uma quantidade razoável de elementos que citem outros documentos da base. Para minimizar esse efeito, pode-se utilizar de *funções de transferência* sobre a distribuição da relevância, de forma a ampliar suas diferenças. Esse processo pode envolver o usuário pela definição de tal função de forma interativa.

CAPÍTULO

3

CENTRALIDADE E TÉCNICAS DE PROJEÇÃO MULTIDIMENSIONAL

3.1 Aspectos gerais

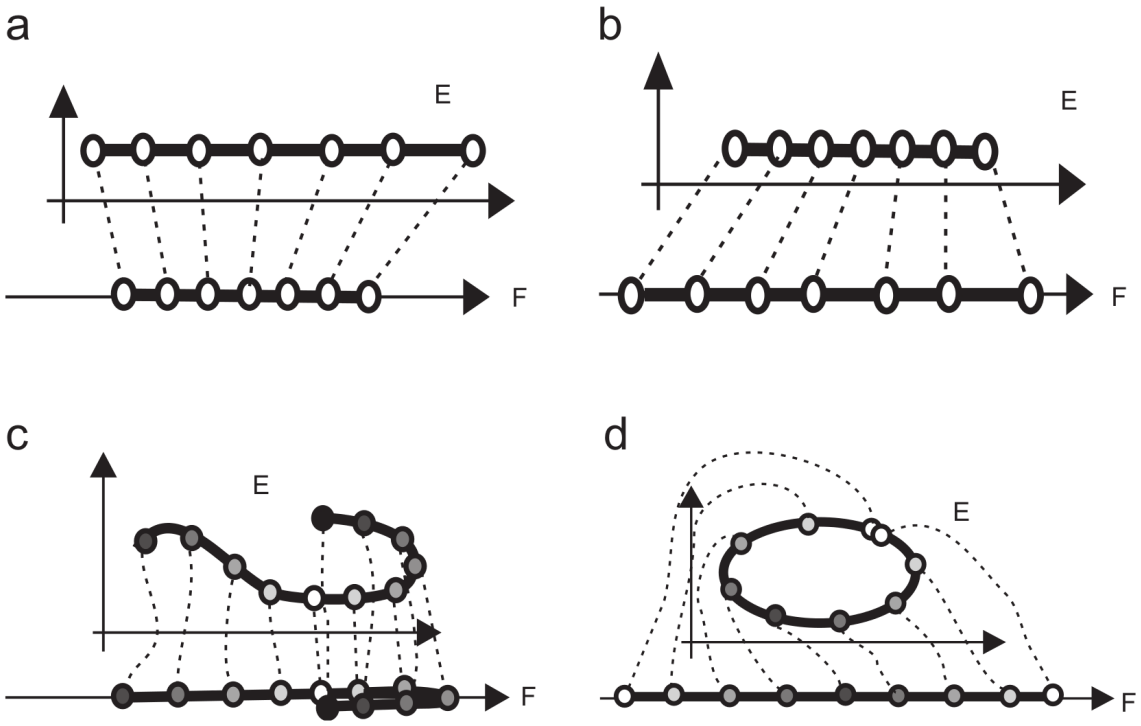
Um dos maiores desafios na área de Visualização de Informação é a busca pelo aumento da confiabilidade na representação de dados multidimensionais (CHEN, 2005). Como citado no Capítulo 1, analisar o resultado obtido por duas projeções diferentes do mesmo dado pode tornar-se uma tarefa complicada sem a ajuda de um critério específico. Em casos mais gerais é comum que os dados analisados não possibilitem explorar metáforas visuais como a descrita no Capítulo 2, que, em sua especificidade, possibilita adicionar sumarizações e relevância aos pontos projetados. Nesse contexto, uma série de medidas de qualidade tem sido propostas de forma a avaliar padrões na visualização de dados multi-dimensionais.

Alguns critérios de avaliação comumente utilizados referem-se a habilidade da medida de qualidade em identificar *clusters* e a relação com a percepção humana de gráficos de dispersão (TATU, 2013). Uma abrangente revisão de medidas de qualidade é provida pela tese de Tatu (2013). No entanto, o foco do estudo neste trabalho é diferente, uma vez que o *pipeline* proposto pela autora utiliza medidas de qualidade sob um procedimento automatizado. Nele, o usuário é capaz de direcionar a projeção multidimensional selecionando dimensões de forma a ter os dados projetados de acordo com a medida de qualidade desejada.

Diferentes linhas de investigação avaliam projeções de acordo como algumas quantidades são preservadas após o procedimento de projeção. Aupetit (2007) classifica as distorções como geométricas e topológicas. As geométricas avaliam distorções em função da escala global ou da relativa, pela distância entre os pontos par a par. Já as topológicas, envolvem a análise de mudança topológica através das mudanças nas vizinhanças dos pontos. Na Figura 31, a linha superior ilustra dois exemplos do caso geométrico e na linha inferior dois exemplos do caso

topológico.

Figura 31 – Distorções na projeção propostas por Aupetit (2007), na primeira linha as geométricas e na segunda as topológicas.



Fonte: Aupetit (2007).

Qualidade baseada em distância

Algumas abordagens baseadas em distância não são invariantes à escala (como medidas de *stress* padrão), significando que mesmo projeções idênticas podem indicar medidas de preservação totalmente diferentes. Além disso, mesmo após o procedimento de normalização, através do Z-Score por ex., conjuntos de dados com *outliers* introduzem um viés na análise já que a distância entre *outliers* a *não-outliers* é proporcionalmente grande, fazendo com que as variações de distâncias entre pontos que não são outliers contribuam menos, na medida de desvio de distância, influenciando na preservação da métrica (MARTINS *et al.*, 2014). Aupetit (2007) investiga medidas de distorção de projeções com enfoque em entidades geométricas (por ex., pontos, arestas e triângulos de Delaunay).

Qualidade baseada em vizinhança

Abordagens topológicas normalmente voltam-se para a investigação de quando pontos são vizinhos apenas no espaço de entrada e quando são vizinhos apenas no espaço visual (SCHRECK; LANDESBERGER; BREMM, 2010). Essas abordagens podem definir medidas bastante rigorosas, uma vez que pequenas perturbações na vizinhança de um ponto podem induzir uma

baixa medida de preservação, embora permaneça em vizinhança próxima a ela. O trabalho de Martins, Minghim e Telea (2015) tenta aliviar esse comportamento descontínuo ao propor uma abordagem em multiescala. Além disso, abordagens mais robustas podem ser utilizadas (RIECK; LEITTE, 2015).

Vários trabalhos dedicam-se na literatura a investigar como a relação de vinhança é afetada com a dimensão do espaço e a dimensão intrínseca dos dados. Beyer *et al.* (1999) analisa que ainda em alguns casos onde a dimensão do espaço não é tão alta (isto é, 10 dimensões), consultas de vizinho mais próximo podem ter considerável perda de sentido, como por ex. quando há correlação não nula entre todas as dimensões. HINNEBURG (2000) investiga o comportamento da diferença entre distância mínima e máxima, sob a norma L_p, entre dois pontos quaisquer de conjunto de pontos, mostrando que para $p \geq 3$ essa diferença tende a zero assintoticamente, ao passo que em $p = 2$ e $p = 1$ mantém-se fixa ou aumenta, respectivamente. A partir disso define uma forma de modificação de pesos nas dimensões para melhorar o significado da busca de vizinhança.

Maier, Luxburg e Hein (2008) investigam como a variação dos parâmetros livres (número de vizinhos k ou raio ε) na construção de grafos de vizinhança pode afetar medidas derivadas do grafo. No trabalho de Luxburg e Alamgir (2013) é investigado como grafos sem peso e não-dirigidos de k-vizinhos podem ser utilizados para extrair informações estatísticas relevantes desses dados (por ex., distribuição inerente às amostras), assumindo que os pontos multidimensionais são identicamente e independentemente distribuídos (i.i.d.). Através de exemplos cuidadosamente construídos, mostra-se que é possível fazer estimativas estatísticas razoáveis, mas apenas para grafos densos em exemplos sintéticos e falha em casos práticos.

Grafos de vizinhança têm uma característica importante que é sua esparsidade, pois eventualmente poucas entradas (ou seja, k entradas) de cada linha de sua matriz de adjacência são não-nulas. Além disso, esses grafos são invariantes à escala uniforme na distribuição dos pontos e contêm informações potencialmente importantes tanto locais (por ex., como a relação de similaridade entre os pontos que compartilham arestas) quanto globais (por ex., *clusters*).

Adicionalmente, ao avaliar a distorção por vizinhança, alguns aspectos devem ser considerados:

- **Semântica:** A ausência de natureza semântica entre vizinhanças, no sentido de que vizinhos de pontos relevantes (e.g, *outliers*) não terem distinção de outras vizinhanças quaisquer (vizinhos de pontos centrais);

- **Sensibilidade:** A natureza estrita da comparação (isto é, conjunto de vizinhos contra outro) pode levar a que pequenas variações (ruídos) nos dados altere essa noção de preservação. Ruído que pode ocorrer tanto da natureza de aquisição dos dados quanto de eventual instabilidade numérica no processo de projeção no espaço visual;

- **Regularidade:** Na área de detecção de anomalias a variação de densidade na vizinhança compromete significativamente a performance dos métodos (VARUN; ARINDAM; VIPIN, 2009). A análise de como métricas de qualidade em projeção multidimensional lidam em tal cenário pode revelar o mesmo comportamento.

Em ambos cenários, geométrico e topológico, não há análise explícita sobre o comportamento de *outliers* nos dados. Várias definições para o que são *ouliers* podem ser encontradas na literatura:

> *"Patterns do not conform a well defined notion of normal behaviour." (VARUN; ARINDAM; VIPIN, 2009)*

> *"Points that lie outside of the set of clusters but are also separated from the noise." (AGGARWAL, 2013)*

Com definições amplas, várias estratégias diferentes são propostas na literatura para seu cálculo, dentre elas: análise de valor extremo, variação de densidade, modelos probabilísticos, dentre outros (AGGARWAL, 2013).

Vale citar que Aupetit (2007) propõe, dentre as citadas medidas de distorção geométricas, uma heurística para analisar as similaridades a um ponto de referência. Apesar de semelhante, a nossa análise tem foco em explorar um campo escalar específico definido sobre os dados, a saber a *profundidade de dados* (*data depth*).

A *profundidade* em um conjunto de dados é um campo escalar particularmente interessante, utilizado em estatística de ordem e estatística não-paramétrica multivariada, onde pouco ou até nada é assumido da distribuição intrínseca dos dados (LIU; PARELIUS; SINGH, 1999; SERFLING, 2006). Transmite uma noção de centralidade nos dados que relaciona-se com estimação de *outliers de valor extremo*, podendo ser determinados como os pontos de menor profundidade num conjunto. Esses pontos podem revelar informações relevantes tais como: um comportamento anormal durante a aquisição ou síntese dos dados; a mistura de conjuntos de dados onde cada qual segue uma distribuições diferente; padrões anormais introduzidos enquanto processando o conjunto de dados (por ex., uma projeção multidimensional) (VARUN; ARINDAM; VIPIN, 2009).

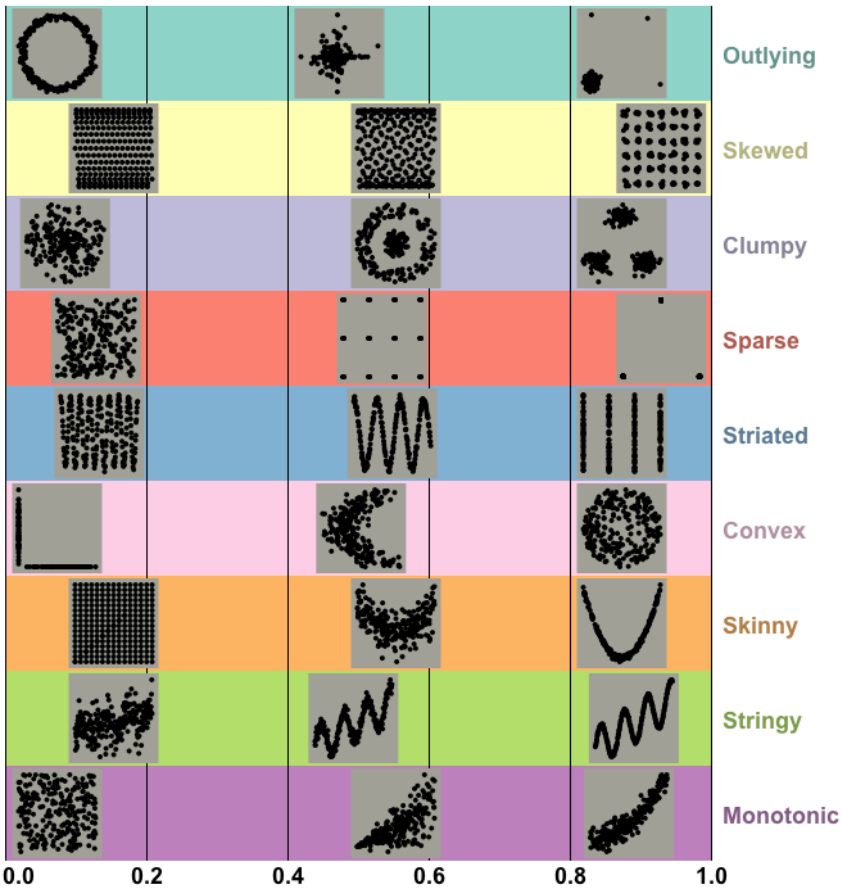
No entanto, mais do que simplesmente conduzir testes estatísticos em distribuições estimadas de dados, informações quantitativas a respeito de quais regiões modificam estimadores estatísticos são requeridas. A combinação dessa noção a possíveis padrões introduzidos pela projeção é também importante nesse contexto.

Uma vez calculada a profundidade dos dados em um certo conjunto, uma análise visual pode prover uma forma qualitativa de avaliar como essa medida foi modificada tanto individualmente para pontos individuais como para suas vizinhanças. Tal informação complementa

eventuais testes estatísticos com a noção de localidade. Cabe notar também que a forma como a noção de centralidade é percebida por quem efetua a análise visual pode ser incorporada no processo, apesar de fora do escopo incial desse trabalho. Essa é uma das limitações que métricas de qualidade possuem, como apontado por Tatu (2013).

Pela classificação de métricas de qualidade proposta por Tatu (2013), a única métrica que possibilita, ao mesmo tempo, análise de gráficos de dispersão, *outliers*, permitindo uma noção de ordenação, consiste no conjunto de métricas proposto por Wilkinson, Anand e Grossman (2005), ilustrado na Figura 32. Utiliza-se de nove diferentes aspectos para qualificar uma projeção, tais como: esparsidade, convexidade, monotonicidade, etc. Apesar de um dos aspectos analisados corresponder a uma quantificação de outliers, ela limita-se ao gráfico de dispersão, impossibilitando a análise comparativa com o dado no espaço original. Além disso, a métrica é global, ou seja, é atribuída ao gráfico de dispersão, não localmente aos seus pontos.

Figura 32 – Aspectos para caracterização de gráficos de dispersão, proposto por Wilkinson, Anand e Grossman (2005).



Fonte: Wilkinson, Anand e Grossman (2005).

3.2 Funções de Profundidade

No caso unidimensional, estatísticas de ordem são significativamente importantes e têm sido utilizadas frequentemente por permitirem o cálculo de medidas estatísticas úteis (como por exemplo, mediana e *outliers*). Sua generalização ao caso multivariado não é imediata uma vez que a generalização individualmente via coordenadas (*componentwise-median*) não define bem uma noção de centralidade dos dados, onde a mediana dos dados pode nem pertencer à combinação convexa dos dados (DING *et al.*, 2007). Uma grande quantidade de abordagens para a mediana multivariada têm sido propostas na área estatística (SMALL, 1990).

Funções de Profundidade (*Depth functions*) fazem-se importantes nesse contexto. Elas podem ser consideradas como uma generalização multivariada de estatísticas de ordem para casos unidimensionais (SERFLING, 2002). Intuitivamente, essas funções definem uma ordenação centro-periferia (*center-outward ordering*) dos dados, permitindo a extração de informações estatísticas relevantes como a mediana multivariada, vista como o ponto mais central. Fornecem também uma forma de relacionar diferentes metodologias estatísticas (como por exemplo, estatística de ordem, estimação de *outliers*) utilizando um único estimador não-paramétrico (SERFLING, 2006).

Embora exista uma intuição simples do que essa noção de centralidade pode representar, as medidas profundidade podem variar de acordo com a função de profundidade utilizada. No decorrer dessa seção será apresentado como essa noção se modifica em alguns exemplos de distribuições simétricas e assimétricas de dados multivariados. Ao longo do texto, serão utilizados de forma indistinta os termos: profundidade, profundidade de dados, centralidade, *data depth*.

Mais formalmente, seja F_n uma distribuição empírica de $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, amostrada de uma distribuição de probabilidade F em $\mathbb{R}^m, m \geq 1$. Uma *função de profundidade* do dado (*depth function*) é uma medida do quão central um dado ponto $\mathbf{x} \in \mathbb{R}^m$ é em relação à distribuição empírica dos dados X.

Como pode ser visto em Zuo e Serfling (2000), Izem, Rafalin e Souvaine (2008), seja $\mathscr{F}$ a classe de distribuições de probabilidade em conjuntos no $\mathbb{R}^m$, F a distribuição de uma variável aleatória em $\mathbb{R}^m$ e seja $D(.,.) : \mathbb{R}^m \times \mathscr{F} \to \mathbb{R}$ uma função limitada, não-negativa, então uma função de profundidade D idealmente deve satisfazer as seguintes propriedades:

P1. **(Invariância afim)** A profundidade de um ponto não deve depender do sistema de coordenadas nem em escalas no dado multidimensional:

$$D(\mathbf{Ax}+\mathbf{b}; F_{\mathbf{Ax}+\mathbf{b}}) = D(\mathbf{x}; F_x),\ \forall \mathbf{x}, \mathbf{b} \in \mathbf{R}^m, A \text{ uma matriz de transformação afim de ordem } m;$$

P2. **(Máximo no centro)** A função de profundidade deve ter valor máximo no ponto mais central, com respeito a noção de simetria associada à distribuição de probabilidade:

$$D(\theta; F) = \sup_{\mathbf{x} \in \mathbb{R}^m} D(\mathbf{x}; F)\ \forall F \in \mathscr{F} \text{ com } \theta \text{ como o ponto mais central;}$$

P3. **(Monotonicidade relativa ao ponto mais central)** A profundidade deve decrescer monotonicamente à medida que o ponto afasta-se do ponto mais central, ao longo de uma direção fixa:

$$D(\mathbf{x};F) \leq D(\theta + \alpha(\mathbf{x}-\theta);F)\,,\ \forall F \in \mathscr{F} \text{ com ponto mais central } \theta \text{ e } \alpha \in [0,1];$$

P4. **(Tender a zero assintoticamente)** A profundidade de um ponto deve tender a zero a medida que sua norma aumenta indefinidamente:

$$D(\mathbf{x};F) \to 0 \text{ quando } \|\mathbf{x}\| \to \infty\,.$$

Em seguida serão descritas algumas funções de profundidade.

Profundidade de Mahalanobis

Função de profundidade de *Mahalanobis* (MHD) é uma das mais antigas propostas. Essa abordagem é largamente utilizada na literatura devido a sua simplicidade e característica intuitiva. É definida como

$$MHD(F_n,\mathbf{x}) = \left[1 + (\mathbf{x}-\mu_{F_n})^T \mathbf{C}_{F_n}^{-1}(\mathbf{x}-\mu_{F_n})\right]^{-1}\,, \tag{3.1}$$

sendo μ_{F_n} a média associada à distribuição empírica F_n dos pontos e $\mathbf{C}_{F_n}$ a matriz de covariância amostral. Faz-se necessário observar que essa função é baseada em uma estimativa não robusta (por exemplo, a média e a covariância) responsável pela imposição de uma séria limitação em situações envolvendo dados contaminados por *outliers*, como no caso ilustrado pela Figura 34.

Profundidade com Fecho Convexo

Entre os diversos estudos a respeito de funções de profundidade presentes na literatura de processamento geométrico e geometria computacional, a profundidade baseada em fecho convexo (CHD) é atrativa devido a sua facilidade de cálculo e entendimento intuitivo. A ideia básica inicia-se pelo cálculo do fecho convexo dos dados de entrada. Todos os pontos pertencentes ao fecho definem a menor profundidade dos dados. Todos esse pontos são então descartados do cálculo e um novo fecho convexo é obtido de maneira a definir um segundo contorno de profundidade. Esse processo é repetido até que a camada mais interior seja encontrada. Embora essa estimativa de profundidade seja satisfatória, como ilustrado pelo Quadro 1, seu uso torna-se inviável em função de um aumento de dimensão.

Profundidade com estatística robusta

A mediana geométrica multivariada, também conhecida como estimador L_1, é a solução teórica do problema de localização de Fermat-Weber. Dadas as amostras X (como anteriormente), e pesos $W = \{w_1, \ldots w_n\}$ para cada amostra, o problema consiste em encontrar um ponto $\mathbf{y}$ que

minimiza a soma com pesos das distâncias entre o ponto $\mathbf{y}$ e as amostras, podendo ser definido como:

$$\mathbf{y} = \arg\min_{\mathbf{x}} \sum_{i=1}^{n} w_i \|\mathbf{x} - \mathbf{x}_i\| \ . \tag{3.2}$$

No caso geral, a solução numérica é estimada pelo uso de um processo iterativo simples, denominado algoritmo de Weiszfeld. Vardi e Zhang (2000) propõem uma solução para esse problema e introduz uma função de profundidade de dados, a $L_1 - Depth$ (L_1D).

Uma vez que $\mathbf{y}$ é a mediana geométrica, ela pode ser considerada o ponto mais central das amostras uma vez que ela minimiza a equação 3.2.

A L_1D é definida como:

$$L_1D(\mathbf{x}) = 1 - \frac{\max\{\mathbf{r}(\mathbf{x}) - \mathbf{w}(\mathbf{x}), 0\}}{\sum_{i=1}^{n} w_i}, \tag{3.3}$$

onde

$$\mathbf{r}(\mathbf{x}) = \left\| \sum_{\mathbf{x}_i \neq \mathbf{x}} w_i \frac{\mathbf{x}_i - \mathbf{x}}{\|\mathbf{x}_i - \mathbf{x}\|} \right\|_2 , \tag{3.4}$$

e

$$\mathbf{w}(\mathbf{x}) = \begin{cases} w_k & \text{se } \mathbf{x} = \mathbf{x}_k, k = 1 \ldots n \\ 0 & \text{caso contrário} \end{cases} . \tag{3.5}$$

Para uma ideia mais intuitiva sobre a Equação 3.3, pode ser considerado o caso onde $w_i = 1$, $\forall i$ e o ponto a ser avaliado a centralidade um $\mathbf{y_i} \notin \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$. Assim, teremos

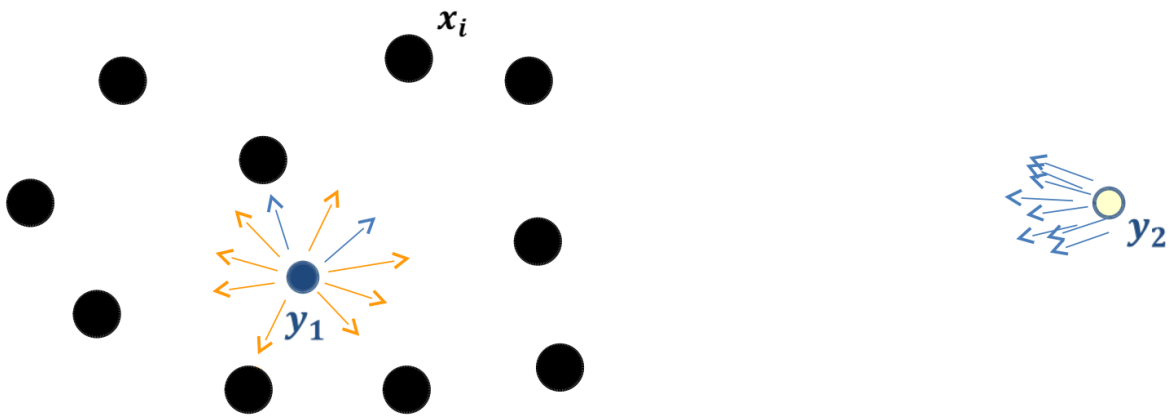
$$L_1D(\mathbf{y_i}) = 1 - \frac{1}{n} \left\| \sum_{j=1}^{n} \frac{\mathbf{x_j} - \mathbf{y_i}}{\|\mathbf{x_j} - \mathbf{y_i}\|} \right\|_2 . \tag{3.6}$$

Analisando a Equação 3.6, fica mais simples de entender o comportamento de um ponto mais central e o de um ponto menos central. Sem perda de generalidade, assume-se $\mathbf{y_1}$ como um ponto de maior centralidade e $\mathbf{y_2}$ como de menor centralidade, como ilustrado na Figura 33. O vetor diferença entre o ponto a ser avaliado e cada ponto do conjunto de dados é normalizado, onde a centralidade dependerá da contribuição do somatório desses vetores de diferença - suas direções. Sendo que a medida de profundidade varia no intervalo $[0, 1]$, onde 1 é o mais central e 0 o menos central.

No caso de pontos centrais como $\mathbf{y_1}$, a quantidade de vetores, de mesma norma, com direções opostas acaba levando ao cancelamento dos termos na soma da Equação 3.6, ilustrados na cor laranja na Figura 33. Com isso, o somatório tem seu valor reduzido e o valor de L_1D tende a um.

No caso oposto, para pontos menos centrais como $\mathbf{y_2}$ na Figura 33, esse processo de cancelamento não ocorre, aumentando o valor do somatório na Equação 3.6, consequentemente fazendo o valor de L_1D tender a zero.

Figura 33 – Ilustração visual da L_1D. Ponto y_1 com alta centralidade e ponto y_2 com baixa centralidade.



Fonte: Adaptada de Ding *et al.* (2007).

Na Figura 34, *outliers* são introduzidos de forma a investigar como os mesmos afetam distribuições de profundidade para estimadores considerados não robustos (como por exemplo o MHD) se comparados a um estimador robusto (ou seja, L_1D). Resultados observados mostram que mesmo após a adição de 20% de outliers, a distribuição de profundidade utilizando L_1D manteve-se praticamente inalterada. Adicionalmente, CHD também apresenta bons resultados na presença de *outliers*. O valor da centralidade varia de zero, ponto menos central a um, ponto mais central.

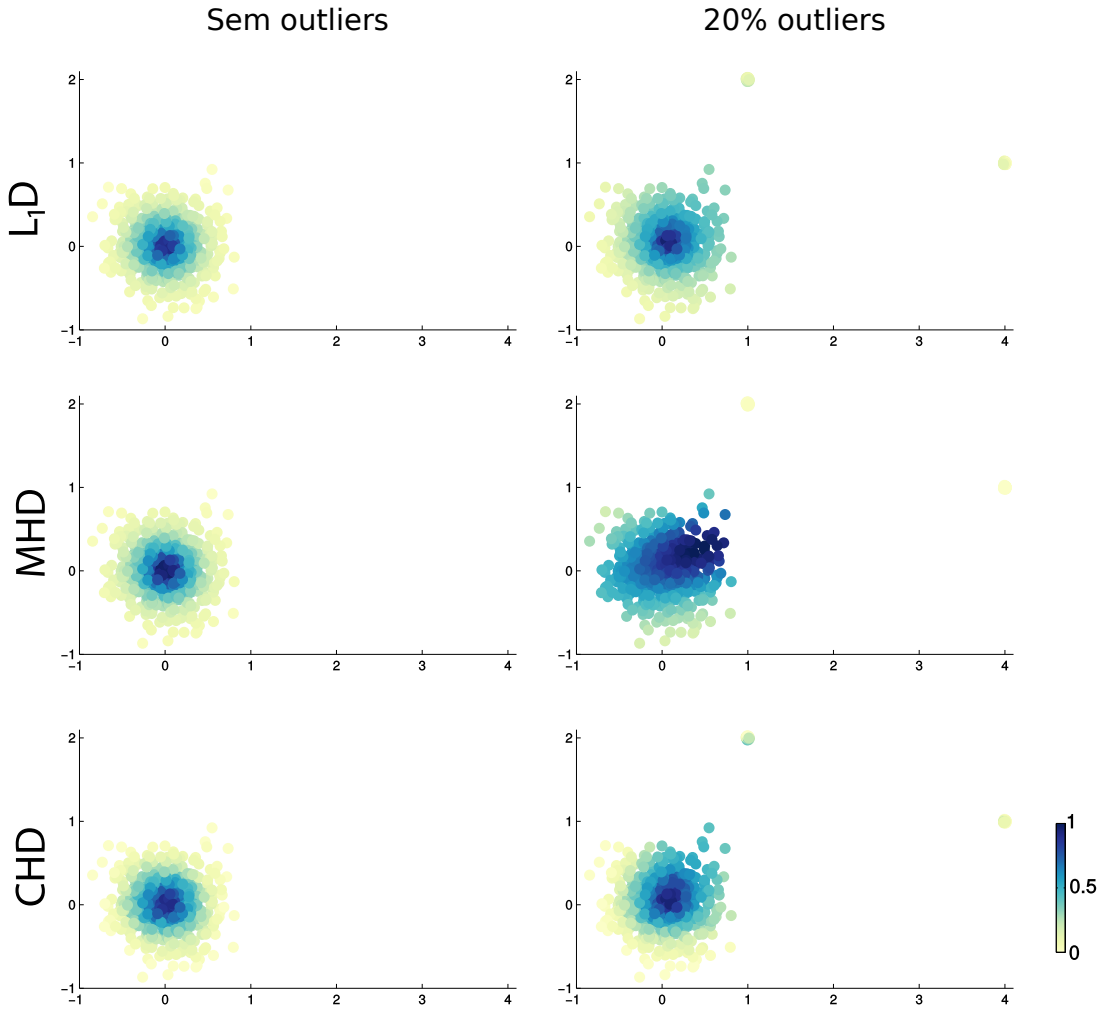
3.2.1 Profundidade em dados multidimensionais

Torna-se importante perceber que a complexidade no cálculo de uma função de profundidade pode impor uma séria limitação em análises envolvendo grande dimensão. Adicionalmente, essas funções acabam sofrendo o problema de mal da dimensionalidade (*curse of dimensionality*). O cálculo do fecho convexo torna-se proibitivo, por ex., já em conjunto de dados com 200 pontos em um espaço de dimensão dez.

Além disso, para a maioria das bases de dados analisadas, a matriz de covariância estimada pela medida de profundidade de Mahalanobis mostrou-se mal-condicionada ou mesmo não invertível, requerendo assim um procedimento de regularização (WON *et al.*, 2013). Há também funções de profundidade baseadas em teoria de *Kernels*, que serão descritos no Capítulo 4. Percebe-se também que embora escalem bem com a dimensão dos dados já que sua complexidade depende apenas do número de instâncias, como listado no Quadro 1, acabam dependendo de ajustes de parâmetros, o que pode dificultar sua utilização para *datasets* multidimensionais e demanda análises posteriores.

Neste Capítulo, o foco das análises é o uso da profundidade L_1. A escolha é justificada uma vez que a função preserva localmente a distribuição de profundidade, mesmo se contaminada por um grande número de *outliers*, como apresentado na Figura 34. Além disso, sua complexidade

Figura 34 – Avaliação de robustez na presença de outliers.



Fonte: Elaborada pelo autor.

permite a análise de conjuntos de dados de dimensões mais elevadas.

Outras quantidades estatísticas podem ser derivadas a partir das estimativas de profundidade de dados (por exemplo *kurtosis, heavy tailedness*), embora não consistindo foco principal do trabalho. Pode-se observar que ao se efetuar a projeção multidimensional, levar em consideração algumas dessas características inerentes a distribuição dos dados pode modificar significativamente o entendimento sobre os dados projetados no espaço visual (LIU; PARELIUS;

Quadro 1 – Algumas características das funções de profundidade analisadas. A dimensão dos dados é representada por m enquanto o número de pontos por n.

Data depth	**Robustez a outliers**	**Complexidade assintótica**
MHD	Não	$O(n+m^3)$
CHD	Não	$O(n\log n + n^{\lfloor \frac{m+(1-m \bmod 2)}{2} \rfloor})$
L_1D	Sim	$O(n^2+nm)$

Fonte: Elaborada pelo autor.

SINGH, 1999; MAATEN; HINTON, 2008).

3.3 Profundidade como uma Medida de Qualidade

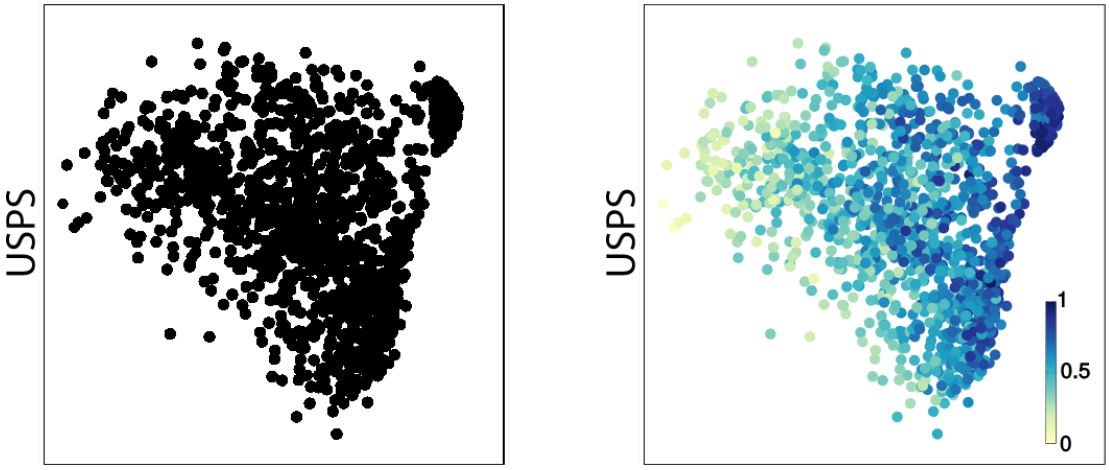
De forma a descrever o uso de profundidade do dado como uma Medida de Qualidade, alguns questionamentos são levantados com o intuito de motivar as escolhas que serão tomadas para análise. Mais especificamente :

Q1: Como visualizar profundidade em conjuntos de dados multidimensionais?

É possível encontrar na literatura uma grande quantidade de trabalhos que fazem uso de *depth functions* para visualizar entidades em espaços bi e tridimensionais (por exemplo na visualização de *ensembles*) (MIRZARGAR; WHITAKER; KIRBY, 2014). No entanto, não há conhecimento de qualquer metáfora visual destinada especificamente à visualização de *depth functions* definidas sobre pontos em um espaço multidimensional com dimensão maior que três.

A forma escolhida nesse trabalho para tratar esse problema se dá por meio de uma abordagem bastante simples, que consiste em efetuar a colorização (*colorcoding*) dos pontos projetados no espaço visual de acordo com o campo escalar de centralidade D^m, calculado nos dados que estão no espaço original, como pode ser visto na Figura 35.

Figura 35 – Análise visual do impacto de utilização do canal de cor para ilustrar a profundidade do dado no espaço original, na projeção do conjunto de dados USPS utilizando PCA.



(a) Sem utilização do canal de cor. (b) Com utilização do canal de cor.

Fonte: Elaborada pelo autor.

Emprega-se uma paleta de cores contínua com três diferentes matizes, do ColorBrewer (HARROWER; BREWER, 2011), de forma a transmitir uma ideia de continuidade da profundidade e enfatizando não somente valores extremos (ou seja, dadas com maior e menor

profundidade) mas também valores intermediários. A ideia aqui descrita é a mesma que a apresentada na Figura 34, embora sendo utilizadas as profundidades calculadas em $\mathbb{R}^m$, o espaço original. O mapeamento de cores varia de azul escuro (ponto mais central, maior profundidade) passando por verde como um ponto de média centralidade para amarelo claro, como de baixa centralidade.

Os experimentos conduzidos pelo uso da estratégia descrita na Seção 3.3.1 e os resultados são apresentados nas Figuras 38, 39, 40 e 41.

Q2: Como as técnicas de projeção multidimensional modificam a centralidade?

Uma vez calculada a profundidade dos dados no espaço original, há possibilidade de se repetir o processo para os dados já projetados no espaço visual, definindo um campo escalar no espaço visual D^2. Dessa forma, tem-se dois campos escalares definidos sobre os dados, ambos no intervalo $[0,1]$.

Uma abordagem simples para investigar como essa noção de profundidade é alterada após uma projeção multidimensional consiste em efetuar a diferença ponto-a-ponto entre ambos os campos escalares, uma vez que há uma relação de um-para-um entre eles. Essa diferença define um novo campo escalar D^d em $[-1,1]$ como:

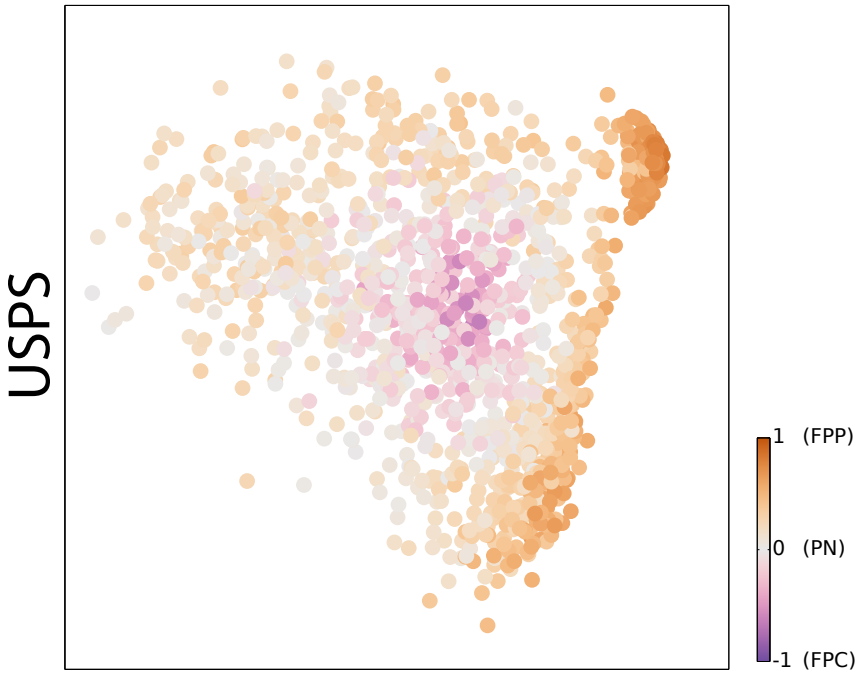
$$D^d(\mathbf{x}) = D^m(\mathbf{x}) - D^2(\mathbf{x}) \ . \tag{3.7}$$

Os valores nesse intervalo possuem uma semântica associada interessante. Ao definir o campo de diferença como na Equação 3.7, seus valores extremos irão transmitir os seguintes comportamentos:

- **Falso ponto periférico (FPP):** O quão próximo o valor for de 1 (ponto laranja), significa que ele era um ponto central (grande profundidade) no espaço original e foi movido para uma região periférica (baixa profundidade) no espaço visual. O nome se dá pois ele é periférico apenas no espaço original;
- **Falso ponto central (FPC):** O quão próximo o valor for de -1 (ponto roxo), implica que um ponto periférico no espaço original foi movido para uma região central (ou seja, grande profundidade) no espaço visual;
- **Ponto Neutro (PN):** O quão próximo o valor for de 0, implica que o ponto não teve sua profundidade modificada, isto é, pontos periféricos no espaço original continuam sendo periféricos no espaço visual, assim como com pontos centrais em ambos os espaços.

Note que essas definições estão bastante relacionadas com a noção de *False-Neighbors* e *Tears* (LESPINATS; AUPETIT, 2011), mas utilizando a profundidade ao invés das vizinhanças.

Figura 36 – Análise visual do impacto de utilização do canal de cor para ilustrar a diferença de profundidade do dado no espaço original e no espaço visual, dado pela Equação 3.7, na projeção do conjunto de dados USPS utilizando PCA.



Fonte: Elaborada pelo autor.

Além disso, está relacionada ao trabalho de Aupetit (2007), que analisa diferentes medidas de distorção porém nenhuma diretamente associada a *outliers*.

Esses três comportamentos são de grande importância, uma vez que podem indicar se possíveis *outliers* no espaço original foram misturados aos dados após a projeção. Até então nenhuma abordagem na literatura lida com essa questão diretamente. De forma a analisar o quão efetiva essa abordagem é, experimentos foram realizados em conjuntos de dados com outliers, descrito na Seção 3.3.1.

Além disso, experimentos foram realizados para avaliar o comportamento de FPP e FPC em dados com características diversas, como: não-gaussianos, alta dimensão, etc. Resultados podem ser observados na Figura 42.

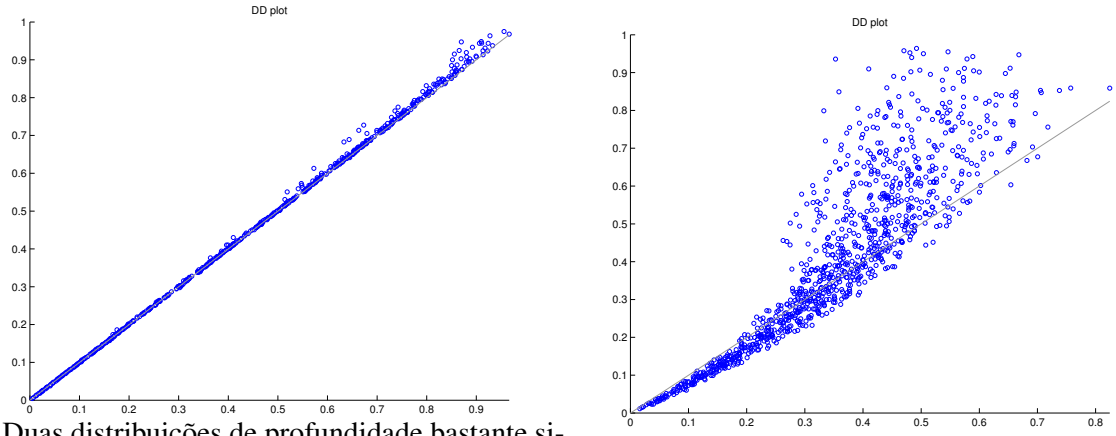
Q3: Como encontrar bons candidatos para direcionar uma projeção multidimensional?

Uma vez que os valores de profundidade dos dados foram calculados, eles podem ser utilizados para um processo de seleção de um subconjunto dos dados no espaço original, através da distribuição dos valores de profundidade. Esse processo é utilizado para definir pontos de controle com uma semântica associada (isto é, centralidade), que permite guiar técnicas de projeção multidimensional através da definição e posicionamento de pontos de controle. Na Seção 3.4 são propostas algumas estratégias com esse objetivo.

3.3.1 Experimentos: Avaliação qualitativa

DD-plots têm sido utilizados por estatísticos para comparar uma distribuição de dados contra outra (LIU; PARELIUS; SINGH, 1999). É definido por um gráfico de dispersão bidimensional onde cada coordenada do ponto é o seu valor de profundidade em uma das distribuições. Isso implica que para duas distribuições idênticas, todos os pontos de um DD-plot ficam sobre a linha $y = x$, de forma similar à comparação visual de distribuições da função de *stress* visto em Joia *et al.* (2011), e que é largamente utilizado na literatura de visualização de informação. A Figura 37 ilustra um exemplo de um DD-plot.

Figura 37 – Exemplo de um DD-plot.



(a) Duas distribuições de profundidade bastante similares.

(b) Variação nas distribuições de profundidade.

Fonte: Elaborada pelo autor.

Apesar de possibilitar a avaliação de mudanças na profundidade de forma simples, uma vez que consiste em observar onde está o ponto relativo à reta $y = x$, não há uma associação direta com a projeção do dado no espaço visual. Em contraste, a análise qualitativa proposta aqui apoiam-se nas estratégias descritas nas questões Q1 e Q2, nas quais a noção de variação de profundidade está associada ao canal visual de cor diretamente no gráfico de dispersão.

Já que não há técnica de projeção multidimensional capaz de preservar explicitamente um campo escala definido nos dados (por ex., medidas de centralidade), optou-se por uma variação da escolha de técnicas em função das hipóteses consideradas sob os dados.

A primeira técnica escolhida é o *Principal Component Analysis (PCA)* , uma vez que a mesma vem sendo muito frequentemente utilizado tanto pela comunidade acadêmica quanto por analistas de dados (LEWIS; MAATEN; SA, 2012). A técnica baseia-se na descoberta de direções de máxima variação ao longo dos dados. A segunda técnica é o *Sammon mapping* (SAMMON, 1969), a qual minimiza distâncias entre pontos de uma forma não-linear. Já a terceira é *Independent Component Analysis (ICA)* (HYVARINEN, 1999), por ser utilizada em distribuições não-gaussianas dos dados. A última, t-SNE (MAATEN; HINTON, 2008), é baseada em probabilidades definidas sobre os pontos, de forma a minimizar diferenças de distribuições de

probabilidades entre os espaços original e visual. Nessa etapa não foi utilizada nenhuma técnica de projeção multidimensional que dependa de pontos de controle.

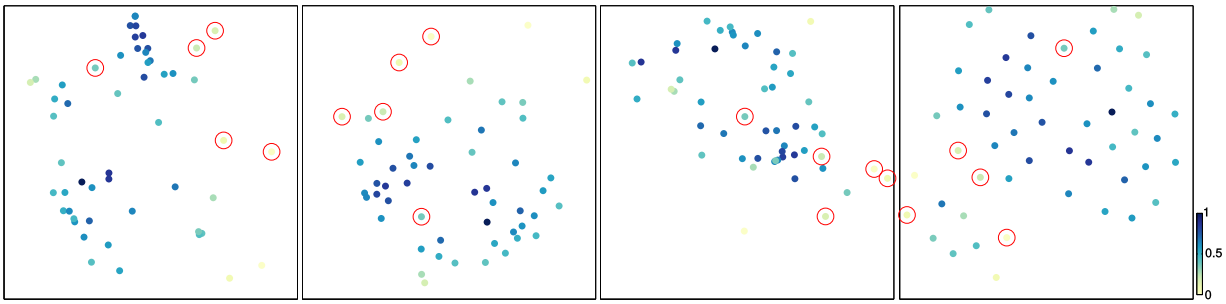
A seguir serão descritos os conjuntos de dados que foram utilizados nos experimentos.

Dados com outliers

Foram conduzidos três experimentos com conjuntos de dados contaminados com *outliers*, que foram marcados por especialistas para avaliar a performance de técnicas de detecção de outliers (CAMPOS *et al.*, 2016). Os conjuntos de dados foram escolhidos de forma crescente de complexidade na detecção de *outliers*: *Parkinson*, *Stamps* e *Hepatitis*, respectivamente. Em todas estes experimentos, os *outliers* foram ilustrados como círculos vermelhos.

O conjunto de dados *Parkinson* é composto por dados médicos de 53 pessoas, cada um com 22 atributos (dimensões). Os cinco pacientes que sofriam do mal de Parksinson são marcados como outliers, o que representa 10% do total. A Figura 38 ilustra o comportamento das técnicas de projeção listadas nesse experimento. É importante notar que em todas as técnicas os *outliers*, circulados em vermelho, tem um relativo baixo valor de profundidade. Mais importante ainda é notar como as técnicas de projeção misturam os outliers com dados centrais.

Figura 38 – Profundidade L_1D calculada no conjunto de dados *Parkinson* com quatro técnicas de projeção multidimensional. Mapeamento de cor de acordo com a profundidade e *ouliers* circulado em vermelho.

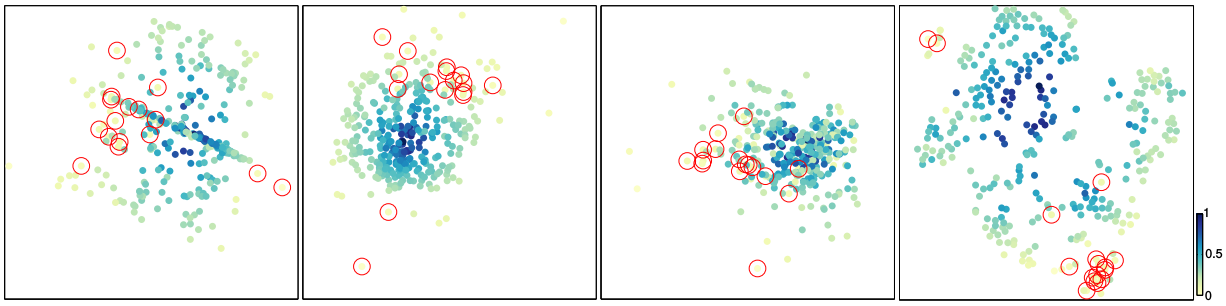


(a) PCA. (b) Sammon. (c) ICA. (d) t-SNE.

Fonte: Elaborada pelo autor.

O segundo conjunto de dados, *Stamps*, é composto por 325 imagens coloridas de carimbos, classificados como genuínos ou modificados (por ex., fotocopiados), onde esses são marcados como *outliers* e são por volta de 5% do conjunto, o que significa 16 *outliers*. Este conjunto de dados é composto por nove características geométricas e de cor, dentre as quais: *bounding box* mínimo, razão de aspecto e densidade de pixel. Na Figura 39 os *outliers* possuem baixo valor de profundidade, de acordo com a L_1D. De forma similar ao experimento anterior, eles também são misturados aos pontos centrais, pelas técnicas de projeção. PCA e ICA revelam os piores resultados nesse cenário.

Figura 39 – Profundidade L_1D calculada no conjunto de dados *Stamps* com quatro técnicas de projeção multidimensional. Mapeamento de cor de acordo com a profundidade e *ouliers* circulado em vermelho.

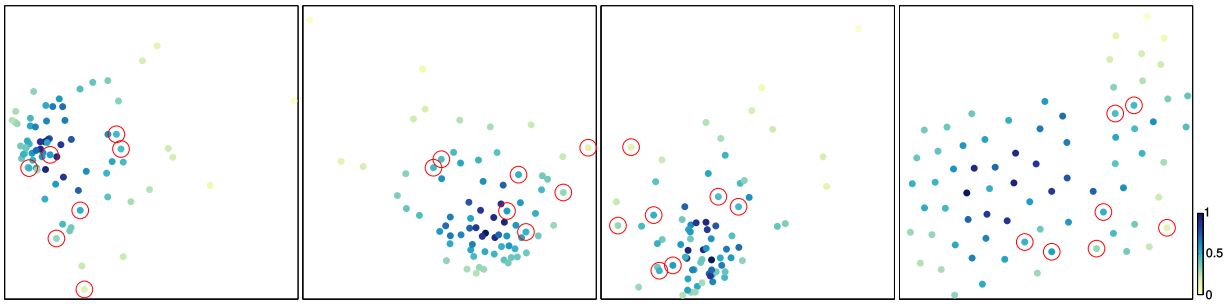


(a) PCA. (b) Sammon. (c) ICA. (d) t-SNE.

Fonte: Elaborada pelo autor.

O terceiro conjunto de dados é composto por 74 pacientes que sofrem de Hepatite (*Hepatitis*), e contém as informações de sobrevivência, *inliers*, ou morte, sendo considerados *outliers*. É composto por 19 atributos e 7 outliers (10%). Dentre os outros conjuntos de dados com outliers testados, esse é o mais complexo para detecção (CAMPOS *et al.*, 2016). Figura 40 reflete tal comportamento, uma vez que os pontos marcados como outliers não apresentam, necessariamente, um baixo valor de profundidade.

Figura 40 – Profundidade L_1D calculada no conjunto de dados *Hepatitis* com quatro técnicas de projeção multidimensional. Mapeamento de cor de acordo com a profundidade e *ouliers* circulado em vermelho.



(a) PCA. (b) Sammon. (c) ICA. (d) t-SNE.

Fonte: Elaborada pelo autor.

Foram realizados outros experimentos visando investigar como as técnicas de projeção lidam com diferentes tipos de conjuntos de dados, os quais regiões menos centrais poderiam, eventualmente, conter *outliers*, ainda que não haja um valor de referência (*ground truth*) como nesses casos apresentados.

Dados Gaussianos

O dataset artificial *AD10* ilustrado na primeira linha da Figura 41 é construído da seguinte forma: quinze *clusters* são gerados seguindo uma distribuição normal $\mathcal{N}(0,1)$, e são

posicionados em vértices aleatoriamente escolhidos de um hipercubo em um espaço de dimensão dez.

A intuição principal deste experimento é verificar como *data depth* e projeções multidimensionais se comportam em um dado com vários *clusters* diferentes, onde cada um deles tem um comportamento conhecido (ou seja, distribuição Gaussiana).

A técnica t-SNE é projetada para obter, também, uma maior separação inter-*clusters*, afetando a preservação de compacidade. Isso possibilita apresentar bons resultados na distribuição de clusters no espaço visual, porém a distribuição intra-clusters acaba sendo apenas parcialmente preservada, o que leva a modificação na noção de distribuição de centralidade. Ao passo que *Sammon mapping*, nos experimentos realizados, ocorre de ser a técnica que melhor preserva a centralidade em comparação com as outras técnicas citadas. Essa análise visual é confirmada por meio de uma avaliação quantitativa, apresentada na Tabela 1.

Além disso, como pode ser notado no experimento ilustrado na Figura 39, ambas mantêm os *outliers* como pontos periféricos na projeção, o que é um bom resultado, se comparado ao PCA e ICA.

Dados Não-Gaussianos

Embora seja comum a suposição de que a distribuição do conjunto de dados segue uma distribuição gaussiana, isso não necessariamente é verdade. Exemplos de distribuições assimétricas (*skewed*) como distribuições log-normal, podem ser encontradas em diversos campos da ciência como: na geologia, medicina humana, microbiologia, ciências atmosféricas, ciências sociais e economia (LIMPERT; STAHEL; ABBT, 2001).

De forma a avaliar o comportamento de distribuições de dados não-gaussianas, criou-se um conjunto de dados seguindo a distribuição log-normal $\ln \mathscr{N}(2000, 0.7)$, em um espaço de dimensão cinco.

Como pode ser observado na segunda linha da Figura 41, uma vez que ICA é adequada para esse tipo de dados (ou seja, não-gaussiano), é esperado que a técnica apresentasse um desempenho melhor que as demais técnicas consideradas. Apesar dos resultados obtidos pelo ICA e pelo Sammon Mapping serem similares, essa ainda apresenta os melhores resultados de preservação de profundidade, visto que em sua região mais central não há pontos de baixa centralidade, como no resultado obtido pelo ICA. Essa análise visual também é confirmada quantitativamente na Tabela 1.

Dados de sensores

O conjunto de dados *ionosphere* consiste em medidas acerca de elétrons livres na Ionosfera terrestre capturados por um radar específico. Ele é composto por dois valores por pulso de sinais eletromagnéticos processados, com 17 pulsos, o que leva a um conjunto de

dados de dimensão 34. Sua importância se dá na busca por alguma evidência de estruturas no sinal (LICHMAN, 2013).

Como pode ser visto na Figura 41, o comportamento da profundidade nesse experimento é praticamente o mesmo em todas as técnicas, com pontos mais centrais do espaço original sendo distribuidos de forma mais central e pontos periféricos sendo distribuídos afastando-se dos centrais. As técnicas *Sammon mapping* e t-SNE produzindo os melhores resultados quantitativos, confirmados pela Tabela 1.

Ao efetuar uma análise visual no resultado obtido pelo Sammon Mapping, é possível perceber que, de forma geral, há pouca mistura de pontos com profundidades bem diferentes. Entretanto, é possível ver alguns casos isolados onde pontos de alta profundidade no espaço original são projetados na periferia.

Dados de alta-dimensão ($m \geq 100$)

O primeiro experimento de dados em alta dimensão, ilustrado na Figura 41, é o *dataset* de dígitos do serviço postal norte-americano (USPS). A base é composta por imagens de dígitos de dez diferentes classes, de zero a nove.

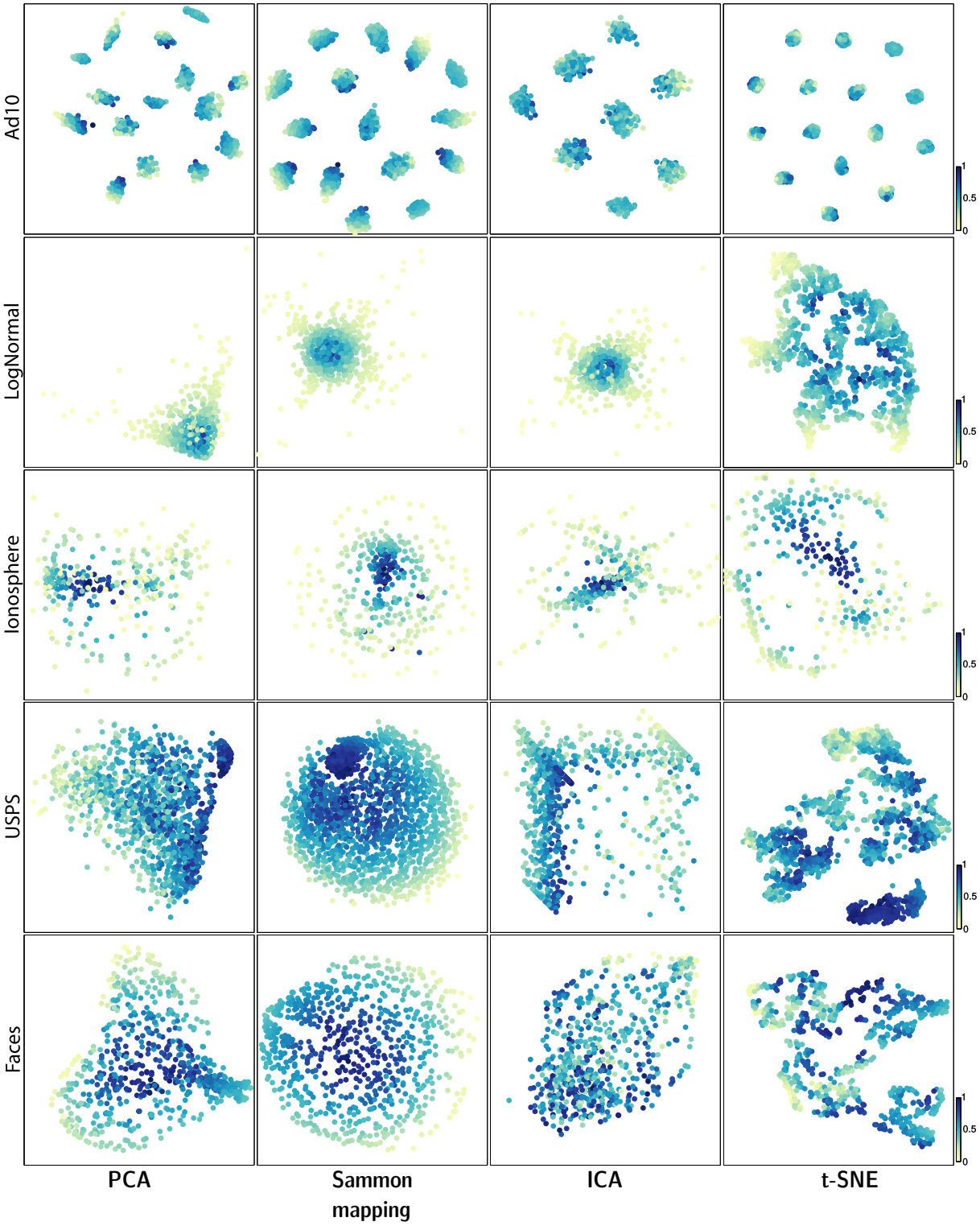
No segundo experimento, ilustrado na Figura 41, o conjunto de dados é composto por 697 instâncias de imagens, de tamanho 64x64 cada, em tons de cinza, de faces em diferentes posições e condições de iluminação (TENENBAUM, 2000). Cada imagem é representada vetorialmente definindo-se pontos em um espaço de dimensão 4096. Tenenbaum (2000) argumenta um outro aspecto interessante sobre esse conjunto de dados, onde há uma conjectura que tais pontos estão distribuídos sobre uma variedade.

Embora nesse experimento o PCA preservou a noção de centralidade, a técnica Sammon mapping produziu os melhores resultados se comparada com as demais técnicas investigadas. A projeção gerada pela t-SNE ainda preserva noção de variação contínua de profundidade definida nos dados de entrada, o que claramente não é observado no caso da ICA, que mistura, no espaço visual, pontos centrais e pontos periféricos do espaço original.

3.3.2 Redução de obstrução (Cluttering)

De forma a evitar poluição visual na visualização que é vista nas Figuras 41 e 42, uma estratégia similar a apresentada por Aupetit (2007) é utilizada. Por meio da Figura 43, regiões com uma alta densidade de pontos produzem células de Voronoi com área menor. Mesmo assim, considera-se que a técnica continua adequada ao processo de análise da preservação de profundidade de dados, uma vez que descontinuidades de cores presentes nas pequenas células de Voronoi indicam pontos com medidas de centralidade no espaço bastante diferentes no espaço original. O fato pode ser claramente observado no lado direito da Figura 43a. Reciprocamente, a distribuição contínua de cores para células de Voronoi vizinhas transmitem a qualidade da

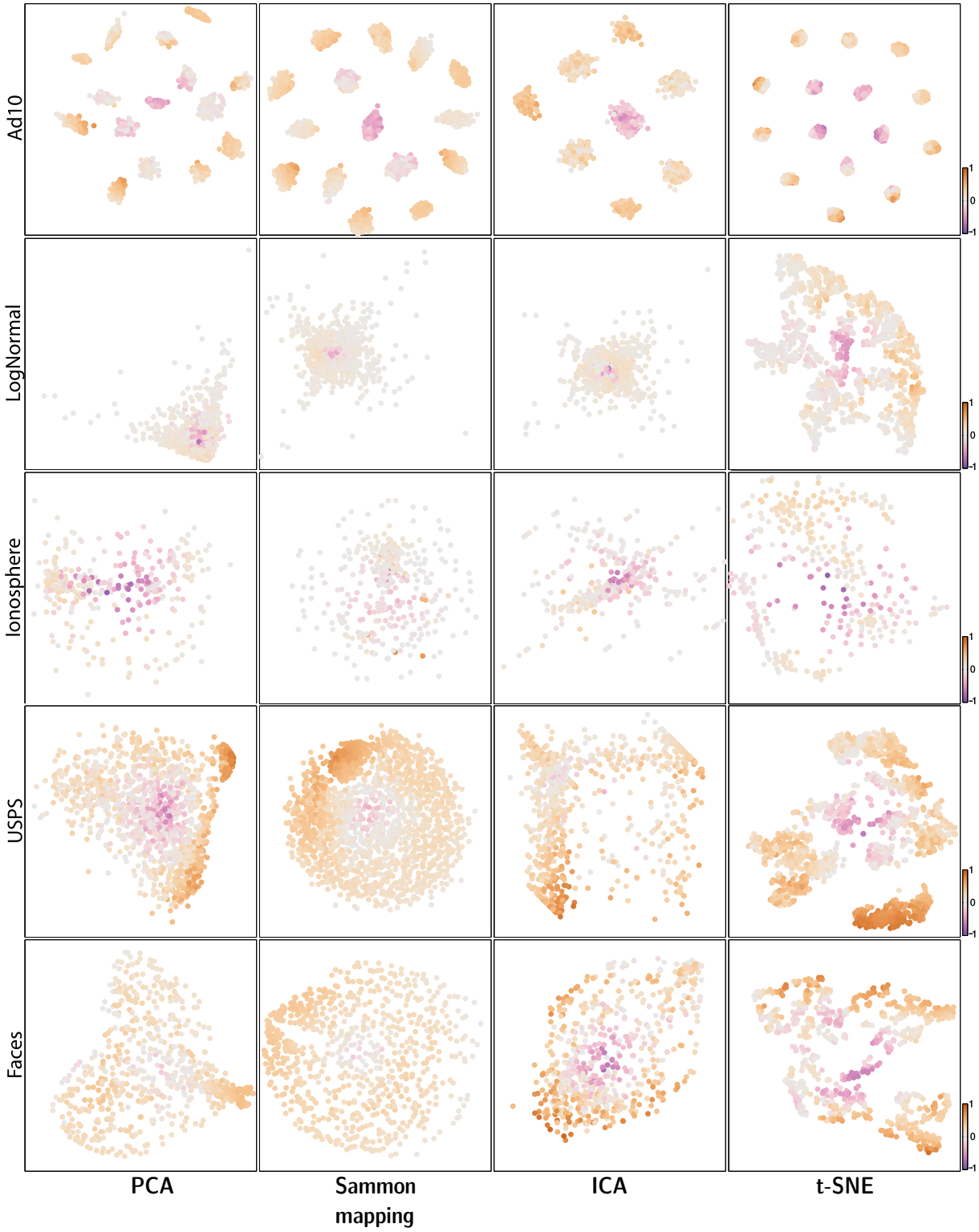
Figura 41 – Profundidade L_1 obtidas em cinco conjuntos de dados e quatro diferentes técnicas de projeção multidimensional.



projeção na preservação da distribuição da profundidade de dados, como pode ser observado na Figura 43b.

O comportamento de FPP e FCP por regiões ao invés de ponto-a-ponto, como demonstrado nas Figuras 43c e 43d, pode beneficiar também projeções com poucos pontos.

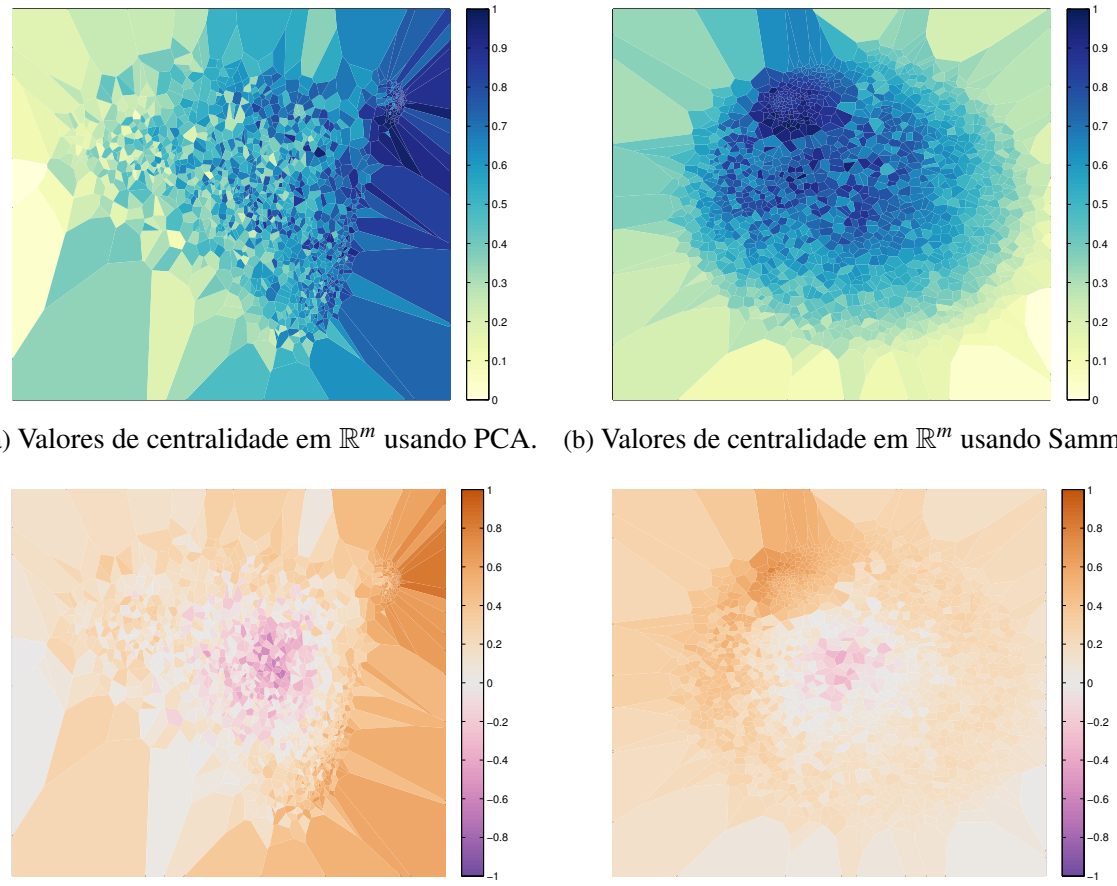
Figura 42 – Diferenças obtidas com a profundidade L_1 em cinco conjuntos de dados diferentes e quatro técnicas de projeção multidimensional. Falsos pontos centrais e falsos pontos periféricos são observados em quase todos os experimentos.



3.3.3 Experimentos: Avaliação quantitativa

O campo escalar de diferença D^d, definido para inspecionar visualmente o comportamento da profundidade dos dados define, de forma imediata, uma medida de preservação de

Figura 43 – Diagramas de Voronoi de projeções do dataset USPS.



(a) Valores de centralidade em $\mathbb{R}^m$ usando PCA. (b) Valores de centralidade em $\mathbb{R}^m$ usando Sammon.

(c) Diferença de valores de centralidade usando PCA. (d) Diferença de valores de centralidade usando Sammon.

Fonte: Elaborada pelo autor.

profundidade. Um vetor $\mathbf{s}$ de n entradas é construído, definido pelo campo escalar de diferença D^d avaliado em cada ponto $\mathbf{x}_i$ do conjunto de dados, ou seja:

$$\mathbf{s} = \left(D^d(\mathbf{x}_1),\ D^d(\mathbf{x}_2),\ \ldots,\ D^d(\mathbf{x}_n)\right). \tag{3.8}$$

Assim é possível definir a *distorção da profundidade de dados* como a norma L_2 do vetor $\mathbf{s}$, tal que quanto mais próximo de zero, menor a distorção. A Tabela 1 quantifica a distorção na profundidade de dados para os vários experimentos efetuados. Faz-se interessante notar a característica de baixa distorção alcançada pelo *Sammon mapping* em todos os experimentos.

Na Figura 44a é ilustrado o quanto cada técnica varia, considerando os valores tabelados de preservação de profundidade de todos os experimentos realizados, mostrando que, de fato, *Sammon mapping* produziu a maior preservação de profundidade. Essa diferença fica bem maior ao considerar apenas os experimentos com *outliers*, ilustrado na Figura 44b.

Já na Tabela 2, são avaliados os valores médios e variâncias da distorção introduzida na profundidade de dados pelas diferentes técnicas, para cada conjunto de dados normalizados pelo seu tamanho. O conjunto de dados *Ad10* como sendo o mais constante entre todas as técnicas

Tabela 1 – Distorção de profundidade de dados medida por D^d. Quanto menor o valor maior a preservação da profundidade pela técnica de projeção multidimensional (melhores resultados em negrido). Linhas estão agrupadas de acordo com as características de cada conjunto de dados.

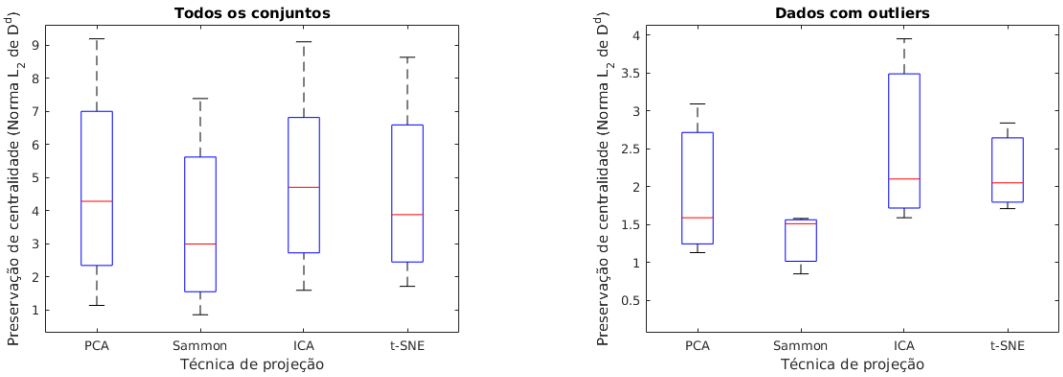
Conjunto de dados (DS)	n	m	PCA	Sammon	ICA	t-SNE
DS1: Parkinson	53	22	1.59	**1.51**	2.10	2.05
DS2: Stamps	325	9	3.09	**1.58**	3.95	2.84
DS3: Hepatitis	74	19	1.13	**0.85**	1.59	1.71
DS4: Ad10	1499	10	8.22	**7.39**	7.81	7.71
DS5: LogNormal	999	5	5.76	**4.22**	5.46	4.59
DS6: Ionosphere	349	34	3.62	**1.76**	3.34	3.17
DS7: USPS	1457	256	9.20	**6.89**	9.11	8.64
DS8: Faces	697	4096	4.94	**4.35**	5.83	5.48

Fonte: Dados da pesquisa.

de projeção e o *Parkinson* como o de maior variabilidade. Os conjuntos de dados com *outliers* representaram os que mais tiveram os valores de profundidade de dados distorcidos durante a projeção.

Um aspecto a ser mencionado é que a alta dimensionalidade não constituiu por si só, nos experimentos, um fator limitador, visto que os resultados obtidos pelo *Faces*, de dimensão 4096, foram semelhantes, considerando a distorção de centralidade, aos obtidos pelo *LogNormal*, de dimensão 5 apenas.

Figura 44 – *Box plots* das distorções de profundidade obtidas pelas diferentes técnicas de projeção.



(a) Todos os conjuntos de dados. (b) Apenas os três conjuntos de dados com outliers.

Fonte: Dados da pesquisa.

Os experimentos foram implementados em MATLAB® e foram executados em um Intel Core i5-5200 CPU com 8 GB of RAM. A Tabela 3 mostra os tempos computacionais para D^m utilizando L_1D.

Tabela 2 – Análise quantitativa média entre distorções introduzidas por diferentes técnicas de projeção multidimensionais, normalizadas pelo tamanho de cada conjunto de dados.

Conjunto de dados (DS)	**Média de distorção** ($\times 10^{-3}$)	**Variância da distorção** ($\times 10^{-3}$)
DS1: Parkinson	34	5.8
DS2: Stamps	8.8	3.0
Ds3: Hepatitis	18	5.4
DS4: Ad10	5.2	0.2
DS5: LogNormal	5.0	0.7
DS6: Ionosphere	8.5	2.4
DS7: USPS	5.8	0.7
DS8: Faces	7.4	0.9

Fonte: Dados da pesquisa.

Tabela 3 – Tempos computacionais em segundos para o cálculo do campo escalar D^m utilizando L_1D.

Dataset	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8
Time (s)	0.009	0.14	0.013	2.78	1.23	0.17	6.41	11.06

Fonte: Dados da pesquisa.

3.3.4 Comparação com CheckViz (LESPINATS; AUPETIT, 2011)

Em geral, como mostrado na Figura 45, os resultados obtidos podem ser bem interessantes ao serem comparados com a técnica de distorção de vizinhança CheckViz (LESPINATS; AUPETIT, 2011), tais como regiões de alta distorção da abordagem proposta também apresentam alta distorção no CheckViz.

O comportamento visual esperado é que pontos periféricos do espaço original (eventualmente *outliers*) sejam mantidos em regiões periféricas após a projeção, definindo células de Voronoi brancas; ou que eles sejam movidos para regiões centrais, definindo células de Voronoi azuis, o que define tais células como possíveis candidatas a conter *outliers*, confirmado nas Figuras 45a e 45b.

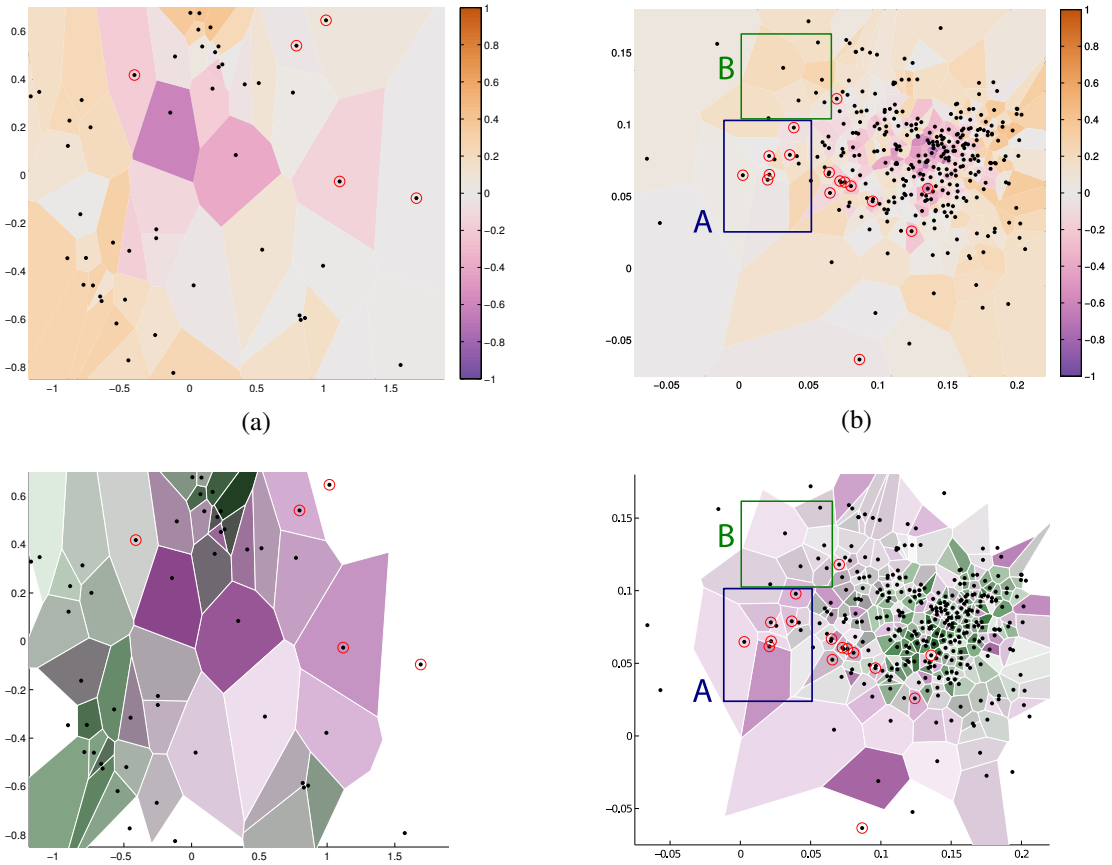
Entretanto, o mesmo não é verificado em algumas células da Figura 46a, que implica que a profundidade calculada no espaço original não atribuiu valores baixos para tais pontos, o que seria esperado para um outlier. Deve-se notar que isso não é uma limitação da metáfora proposta mas sim da capacidade da função de profundidade detectar *outliers* adequadamente, o que aparentemente não ocorreu nesse exemplo utilizando a L_1D, comportamento que deve ser melhor investigado.

Por outro lado, o comportamento esperado da CheckViz consiste em que candidatos a outlier ou sejam mantidos em regiões periféricas, também definindo células de Voronoi brancas; ou misturados à regiões mais centrais, definindo células de Voronoi roxas, os falsos vizinhos, que são pontos que são vizinhos apenas no espaço visual. De forma análoga à nossa abordagem, ambos tipos de células podem conter outliers, comportamento verificado nas Figuras 45d e 46b.

Entretanto, como pode ser visto na Figura 45c outliers também podem definir *Tears*,

pontos que são vizinhos apenas no espaço original, ilustrados como células de Voronoi verdes. Isso pode ocorrer devido a sua necessidade de ajuste de parâmetros. Durante os experimentos, o parâmetro livre σ da técnica CheckViz foi selecionado de acordo com uma heurística proposta pelos autores, a saber: é escolhido como a distância média entre cada ponto e seu quinto vizinho mais próximo no espaço original.

Figura 45 – Abordagem proposta (linha de cima) e CheckViz (linha de baixo).

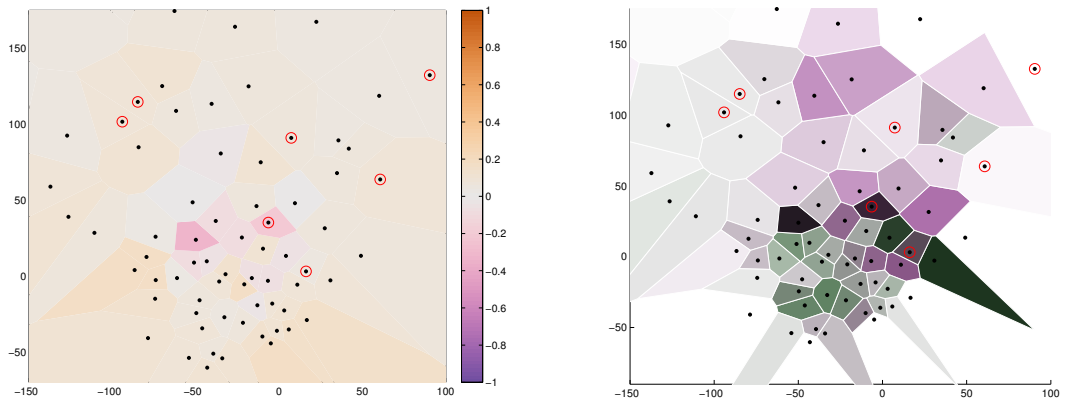


Fonte: Elaborada pelo autor.

Outro aspecto interessante pode ser visto nas Figuras 45b e 45d, que ilustram uma distinção importante no tipo de análise efetuada. Mais especificamente, a análise será focada em duas regiões $A : [-0.01, 0.05] \times [0, 0.1]$ and $B : [0, 0.07] \times [0.10, 0.16]$. CheckViz visualmente indica praticamente a mesma distorção para todas as células nas regiões A e B. A abordagem por profundidade revela comportamentos bem diferentes de ambas regiões: a) Região A contém pontos periféricos no espaço original que ainda são periféricos após a projeção (pontos neutros); b) Região B contem FPPs, isto é, falsos pontos periféricos, perderam centralidade após a projeção, de acordo com a estimativa de profundidade.

Essas diferentes possibilidade de comparação produzem impactos no entendimento do efeito da projeção sobre os dados. Por exemplo, ao utilizar a abordagem proposta pode-se

Figura 46 – Abordagem proposta (esquerda) e CheckViz (direita), ambos calculados no conjunto de dados *Hepatitis* usando *Sammon mapping*.



Fonte: Elaborada pelo autor.

identificar regiões candidatas a possuírem outliers buscando por células menos centrais com cores neutras, ou células centrais com cores azuis.

Em projeções onde a percepção de centralidade não é simples de ser avaliada, como em distribuições de pontos em diferentes clusters, ou distribuição de forma não-convexa, faz-se necessário quantificar a diferença da centralidade computada, por alguma função de profundidade, com a centralidade percebida pelo analista no espaço visual. Estudos de caso podem ser feitos para quantificar essa diferença.

3.4 Aplicação: Seleção automática de pontos de controle

O cálculo de funções de profundidade define um campo escalar entre os dados de entrada, em que cada valor descreve uma noção de ordenação centro-periferia. Por meio dessa informação, é possível dirigir a projeção multidimensional através da noção de centralidade. De forma a explorá-la, uma avaliação de diferentes estratégias para escolha de amostras do espaço de entrada e posicionamento de pontos de controle no espaço visual é realizada. Ambos os pontos de alta e baixa centralidade são considerados como amostras.

Algumas técnicas de projeção multidimensional são semi-supervisionadas, ou seja, dependem da seleção de pontos no espaço original e seu posicionamento no espaço visual como restrições para a projeção (isto é, pontos de controle), possibilitando uma forma de guiar a projeção. Assim, possibilita que o processo de projeção seja guiado pelo usuário através da escolha de pontos com alguma semântica associada (por ex., centroide de *clusters*, *outliers*) e por posicioná-los adequadamente no espaço visual, visando que a projeção preserve o *layout* pretendido tanto quanto possível.

Esse processo de seleção e ajuste de pontos de controle é diretamente dependente da

tarefa que deseja-se. Por exemplo, pode-se utilizar a informação de classe dos pontos para posicionar pontos de diferentes classes em regiões distintas do espaço visual, com o objetivo de aumentar a separação interclasse entre os dados.

Uma vez que o cálculo de profundidade define um campo escalar sobre os dados no espaço original, com uma semântica de ordem associada, pode-se utilizar essa informação para guiar a projeção visando preservar a noção de centralidade, de forma que possíveis *outliers* não estejam em regiões mais centrais no plano.

De forma a explorar essa informação, foram implementadas estratégias diferentes de amostragem no espaço original e de posicionamento no espaço visual, comparadas ao posicionamento aleatório de pontos de controle. A ideia de ambas estratégias é preservar a noção de centralidade dos dados através de um posicionamento radial no espaço visual, devido a noção imediata de centralidade em uma distribuição radial dos pontos.

Várias técnicas de projeção objetivam minimizar as distorções de distância entre os espaços original e visual. A utilização da informação de centralidade na escolha dos pontos de controle foi avaliada através da medição na distorção na centralidade, como descrito na Seção 3.3.3, bem como da distorção entre as distâncias no espaço original e visual, definida como *stress*:

$$stress = \frac{\sum_{ij}(d_{ij} - \bar{d}_{ij})^2}{\sum_{ij} d_{ij}^2}, \tag{3.9}$$

onde d_{ij} e $\bar{d}_{ij}$ representam as distâncias entre os pontos $\mathbf{x_i}$ e $\mathbf{x_j}$, nos espaços original e visual, respectivamente.

Como não há uma técnica de projeção que objetive preservar essa noção de centralidade, foi utilizada a técnica *Local Affine Multidimensional Projection* (LAMP) (JOIA *et al.*, 2011) como a análise mais próxima possível, devido a sua efetividade na preservação de distâncias localmente. Ela é uma técnica de projeção multidimensional semi-supervisionada, ou seja, que permite que o processo seja guiado através de dois aspectos: a) primeiro passo é a seleção de pontos, no espaço original, como *pontos de controle*, que serão usados para guiar a projeção dos outros pontos. A técnica preserva tanto quanto possível a distância a esses pontos; b) O segundo passo é o posicionamento de tais pontos no espaço visual. A técnica é definida em seguida.

Para cada ponto $\mathbf{p}$ no espaço original, o seu mapeamento para o espaço visual é efetuado através de uma transformação afim $f_{\mathbf{p}}(\mathbf{x}) = \mathbf{x}M + \mathbf{t}$ que minimiza o seguinte funcional:

$$\sum_i \alpha_i(\mathbf{p}) \|f_{\mathbf{p}}(\mathbf{x_i}) - \mathbf{y_i}\|^2, \text{ sujeito a } \mathbf{M}^\top \mathbf{M} = \mathbf{I}\ , \tag{3.10}$$

onde $\{\mathbf{x_i}, \mathbf{y_i}\}$ são os pontos de controle no espaço original e no espaço visual respectivamente, e o termo de peso:

$$\alpha_i(\mathbf{p}) = \frac{1}{\|\mathbf{x_i} - \mathbf{p}\|^2}\ . \tag{3.11}$$

Seja k o número de pontos de controle, $\hat{\mathbf{x}}_\mathbf{i} = \mathbf{x}_\mathbf{i} - \tilde{\mathbf{x}}$, $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \tilde{\mathbf{y}}$, com $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ sendo os centroides dos pontos de controle no espaço originale no espaço visual, respectivamente. Por construção, as matrizes A e B são tais que

$$\mathbf{A} = \begin{bmatrix} \sqrt{\alpha_1 \hat{\mathbf{x}}_\mathbf{1}} \\ \vdots \\ \sqrt{\alpha_k \hat{\mathbf{x}}_\mathbf{k}} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \sqrt{\alpha_1 \hat{\mathbf{y}}_\mathbf{1}} \\ \vdots \\ \sqrt{\alpha_k \hat{\mathbf{y}}_\mathbf{k}} \end{bmatrix}, \tag{3.12}$$

a minimização do funcional da Equação 3.10 é calculada de forma exata (*closed-form*) como a solução ao seguinte problema de Procrustes:

$$\mathbf{M} = \mathbf{UV}, \quad \mathbf{A}^\top \mathbf{B} = \mathbf{UDV} \,. \tag{3.13}$$

Com o objetivo de minimizar o efeito da diferença de escala na medida *stress*, pontos de controle foram mapeados para um disco bidimensional de raio $d_{max}/2$, onde d_{max} corresponde à máxima distância entre dois pontos do espaço original. A performance de cada estratégia é mostrada na Tabela 4, comparando suas medidas de *stress* e preservação de profundidade, como norma do campo D^d.

Em todos os experimentos nessa seção os pontos de controle são desenhados como *glyphs* pretos em forma de $\times$.

3.4.1 Amostragem aleatória (AA)

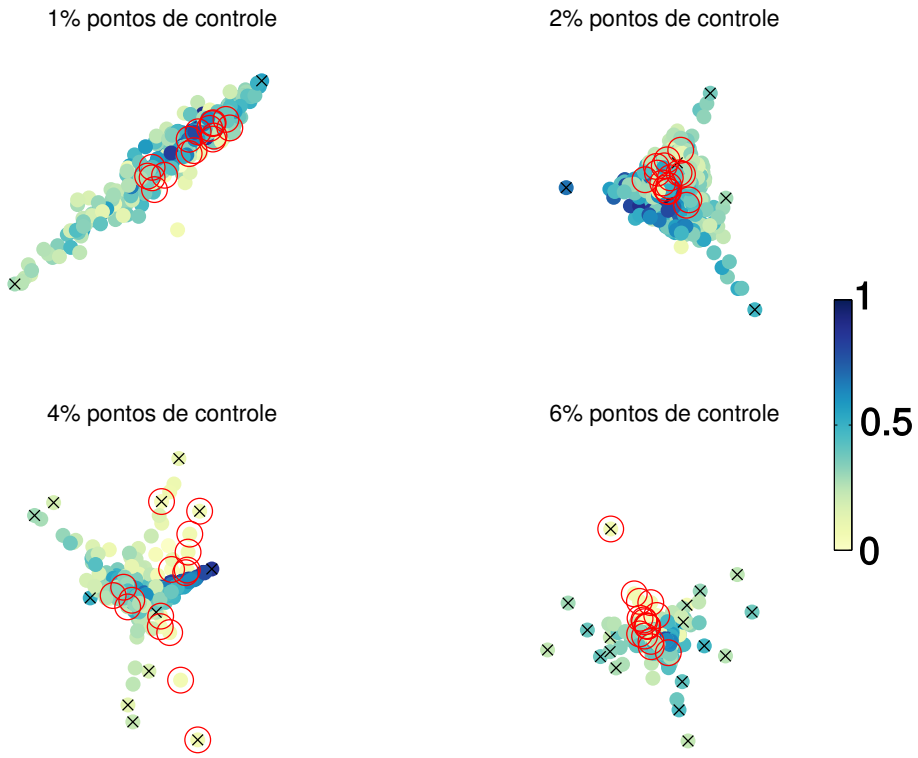
Uma estratégia direta de amostragem consiste em selecionar pontos no espaço original randomicamente e posicioná-los aleatoriamente dentro do disco bidimensional de raio $d_{max}/2$. O uso dessa estratégia é justificado no sentido de se identificar como a técnica LAMP se comportaria caso os pontos de controle fossem escolhidos de forma aleatória, ou seja, sem que informações sobre os dados (por exemplo, profundidade) fossem levadas em consideração.

A Figura 47 ilustra a projeção obtida para a base de dados de carimbos, variando-se o número de pontos de controle.

A natureza aleatória nas escolhas dos pontos de controle e seus posicionamentos não carrega uma semântica mais informativa, especialmente no que se refere à profundidade. Vale ressaltar que mesmo espalhados de forma aleatória no espaço visual, a LAMP, como característica da técnica, garante a preservação local de distâncias das vizinhança dos pontos de controle. O objetivo de sua análise na comparação é por ser a forma mais geral para seleção de pontos de controle.

De forma a considerar profundidade, amostra-se valores calculados dessa medida por meio de duas estratégias: uniforme e não-uniforme, ou seja, adaptada à distribuição de profundidade no espaço original.

Figura 47 – Estratégia de amostragem aleatória em uma base de dados de carimbos. Marcas em forma de x identificam pontos de controle e círculos vermelhos são *outliers*.



Fonte: Elaborada pelo autor.

3.4.2 Amostragem uniforme de profundidade (AUP)

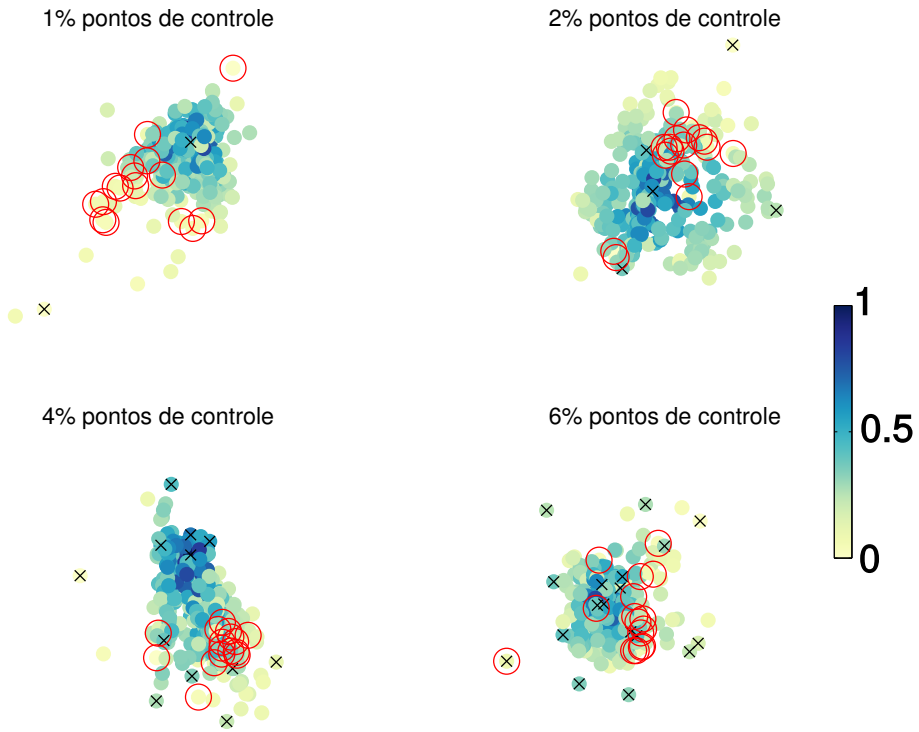
Na primeira abordagem, candidatos a ponto de controle são selecionados no espaço original por uma amostragem uniforme de seus valores de profundidade. Uma vez escolhidos, os mesmos são posicionados no espaço visual da seguinte forma: distribui-se os pontos de controle dentro do mesmo disco bidimensional previamente definido de forma que o raio de cada um desses pontos represente sua profundidade no espaço original. Nesse sentido, quanto maior a profundidade calculada do ponto, menor o raio a ele atribuído.

A formulação adotada utiliza coordenadas polares, por diretamente representar esse comportamento. Para uma posição geral de pontos de controle, o parâmetro θ é gerado aleatoriamente seguindo uma distribuição uniforme no domínio de $[0, 2\pi]$ e o raio ρ é uniformemente amostrado de $\left[0, \frac{d_{max}}{2}\right]$. Esse comportamento é ilustrado na Figura 48. Aqui, faz-se importante perceber que mesmo utilizando apenas 14 pontos de controle ($\sim 4\%$ dos pontos), a projeção se mostrou capaz de manter pontos periféricos e *outliers* na parte inferior separados de pontos mais centrais. Comportamento semelhante de separação de periféricos e centrais pode ser observado para 2% de pontos de controle, porém com uma distribuição mais radial.

Como pode ser observado na Tabela 4, a abordagem aqui adotada melhor preserva *stress* se comparada com a amostragem aleatória. Nota-se também que ela reduz a distorção de profundidade, pela construção radial onde pontos mais centrais do espaço de entrada são

posicionados no centro do disco bidimensional.

Figura 48 – Amostragem da base de dados de carimbos utilizando a estratégia uniforme.



Fonte: Elaborada pelo autor.

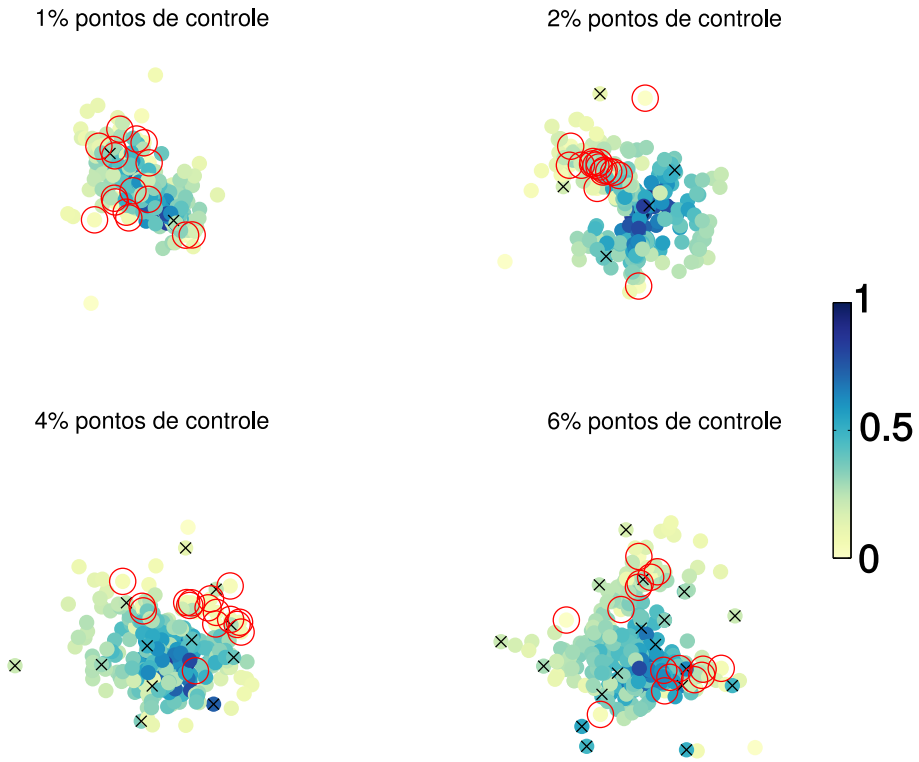
3.4.3 Amostragem não-uniforme de profundidade (ANUP)

O segundo passo no processo de análise da influência da estratégia de amostragem se refere ao uso da informação da distribuição de profundidade dos dados para adaptar o processo de amostragem. Utiliza-se a distribuição de profundidade empírica do espaço original (ou seja, distribuição de D^n) para se efetuar uma seleção aleatória e adaptativa de pontos de controle e para se posicionar adequadamente esses pontos no espaço visual por meio de diferentes raios ρ.

A intuição desse tipo de esquema ponderado está associada a simular a distribuição de profundidade depois da projeção. Na Figura 49, um comparativo dessa junto à abordagem uniforme é apresentado.

Como pode ser observado na Tabela 4, a abordagem adotada é novamente superior à amostragem aleatória em *stress* e na preservação de profundidade, como esperado. No entanto, se comparada com AUP, melhorias não aparentam ser estatisticamente significantes. Porém, ao observar os resultados ilustrados nas Figuras 48 e 49, num comparativo entre as abordagens pode-se argumentar pela identificação de distribuições levemente mais regulares para valores de profundidade intermediários geradas pelo uso da abordagem adaptativa, para todos os números de pontos de controle testados.

Figura 49 – Amostragem não-uniforme do conjunto de dados *Stamps*.



Fonte: Elaborada pelo autor.

Tabela 4 – Análise quantitativa de diferentes estratégias de amostragem (melhores resultados em negrito).

Dataset	AA	AUP	ANUP
Parkinson (stress)	3.42 (±2.55)	0.31 (±0.04)	**0.27** (±0.06)
Parkinson (distorção D^d)	2.34 (±0.18)	1.88 (±0.05)	**1.81** (±0.23)
Stamps (stress)	2.68 (±2.1)	0.26 (±0.05)	**0.21** (±0.04)
Stamps (distorção D^d)	5.70 (±0.62)	3.79 (±0.53)	**3.16** (±0.21)
Hepatitis (stress)	3.79 (±1.8)	0.22 (±0.06)	**0.14** (±0.05)
Hepatitis (distorção D^d)	2.62 (±0.30)	1.94 (±0.31)	**1.87** (±0.13)

Fonte: Dados da pesquisa.

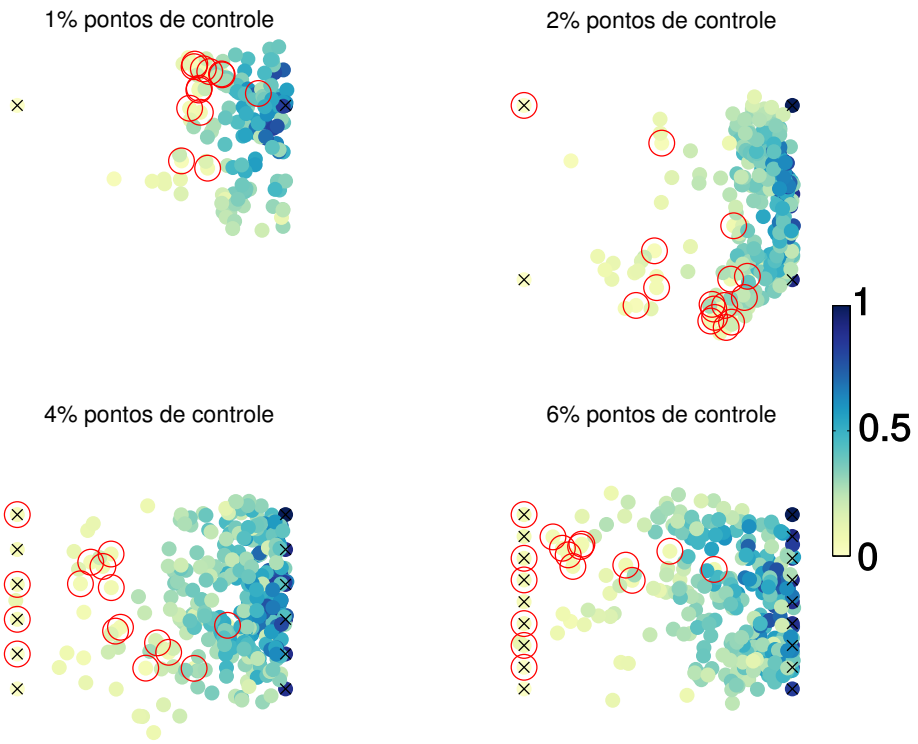
3.4.4 Posicionamento de pontos baseado em tarefas específicas

Mesmo que as estratégias de amostragem uniforme (AUP) e não-uniforme (ANUP) objetivem o uso da medida de profundidade e sua preservação no processo de projeção, estratégias de amostragem permitem que *layout* de pontos de controle seja orientado a outras tarefas específicas, como detalhado a seguir.

No experimento ilustrado na Figura 50, pontos de controle com as menores medidas de profundidade são posicionados no lado esquerdo do espaço visual, enquanto os de maior profundidade são posicionados no lado direito.

Observa-se que a medida que aumenta-se a quantidade de pontos de controle, os pontos

Figura 50 – Amostragem da base de dados de carimbos tendo valores extremos de profundidade como pontos de controles.



Fonte: Elaborada pelo autor.

que são *outliers* de valor extremo, (isto é, que possuem um baixo valor de profundidade), serão pontos de controle no lado esquerdo, aproximando outros eventuais *outliers*, também de baixo valor de profundidade, em sua direção, mesmo com a pequena quantidade de dez pontos de controle ($\sim 4\%$).

A principal razão de se impor essa separação dos dados em acordo com valores extremos de profundidade se deve a exploração da possibilidade de definição de regiões de projeção imunes à contaminação de *outliers*. Isso torna possível explorar, no espaço visual, conjuntos de dados onde não tenha-se a informação, à priori, sobre e existência ou não de *outliers*, controlando seu posicionamento no espaço visual.

O experimento demonstra como o aumento do número de pontos de controle guiados pela profundidade pode ter impacto nesse processo. A média de *stress* obtida por esse experimento foi 0,21 ($\pm 0,030$) e a média normal da distorção D^d foi de 4,44 ($\pm 0,34$).

3.5 Discussão e Limitações

Utilizando técnicas de projeção não-supervisionadas, para todas os conjuntos de dados utilizados nos experimentos, o *Sammon mapping* mostrou-se como a técnica que melhor preserva a medida de profundidade. Como a noção de profundidade de dados acaba dependendo da noção

de ordem centro-periferia, técnicas que preservam melhor localmente as medidas de similaridade tendem a preservar essa noção de profundidade indiretamente, o que observou-se com o *Sammon mapping* e t-SNE. Entretanto, com o objetivo de aumentar a distância inter-*clusters*, o t-SNE possibilita um maior espalhamento dos dados no espaço visual, o que acaba por distorcer um pouco mais a profundidade calculada nos dados.

Como pôde ser observado na Figura 42, nenhuma das técnicas preservou totalmente a profundidade dos dados após a projeção. Pode ser notado, em particular, que em quase todos os exemplos - a menos dois dos resultados para o conjunto de dados Faces - os pontos mais centrais no espaço visual eram majoritariamente FPC, isto é, falsos pontos centrais, pintados de roxo. A abordagem proposta ilustra diretamente onde esses pontos que tinham maior centralidade residem após a projeção, sendo pintados de laranja.

Em relação as estratégias de amostragem, não foram observadas diferenças estatísticas significativas entre as estratégias uniforme (AUP) e não-uniforme (ANUP). Com isso, o uso da estratégia uniforme pode ser preferido devido à sua maior simplicidade, ainda que a estratégia não-uniforme possua um leve aumento na regularidade de distribuição dos pontos. No entanto, há possibilidade de se definir um outro esquema ponderado de forma que valores próximos aos extremos apresentem os maiores pesos. Esse processo pode reforçar uma separação no plano, similarmente discutido no experimento da Seção 3.4.4, com a diferença de ser realizado de forma a se preservar a distribuição de profundidade.

Ao se analisar diferentes estratégias de amostragem, dois pontos podem ser discutidos:

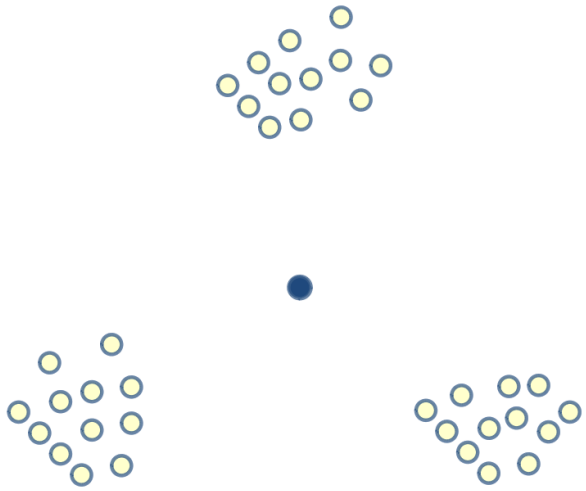
1. A definição do campo de profundidade permite a interação de usuário por meio de diferentes layouts de pontos de controle. O usuário pode fazer uso de ferramentas interativas para indicar em quais regiões valores extremos de profundidade deveriam ser projetados, por exemplo;

2. Embora o *stress* e a preservação da medida de profundidade sejam diferentes em natureza, em todos os experimentos realizados utilizando pontos de controle, ambas as medidas mostraram-se ter um comportamento acoplado, ou seja, embora em diferentes escalas ambas as médias decrescem conjuntamente. Em casos onde o *stress* é reduzido em torno de 90%, a preservação de profundidade acompanha essa redução em torno de 30% para o caso da estratégia de amostragem aleatória.

Um aspecto interessante no processo de avaliação dessas três diferentes estratégias é que, embora LAMP não objetive explicitamente preservar medidas além da distância local na projeção, a preservação da profundidade dos dados pode ser alcançada por meio da exploração dos layouts de pontos de controle. Isso se dá mesmo sendo considerados poucos pontos de controle. As estratégias aqui descritas são experimentos *ad-hoc* realizados nessa direção e que podem ser melhorados por meio de otimização.

Como observado na Figura 40, o uso de L_1D pode produzir resultados pobres ao serem considerados conjuntos de dados complexos, quanto à dificuldade de identificação de *outliers*. Pontos que não sejam *outliers* de valor extremo, como o caso ilustrado na Figura 51 induziriam uma alta centralidade em pontos que são claramente *outliers*, sendo uma limitação dessa análise por funções de profundidade. Porém, ao utilizar uma técnica que preserve a centralidade de tais pontos, os mesmos poderiam ser exibidos como pontos neutros no espaço visual, o que levaria a sua identificação como um possível *outlier* por quem está analisando o gráfico de dispersão, já que destoa do comportamento dos de mais e não é um erro introduzido pela projeção.

Figura 51 – Limitação da utilização de profundidade como esquema geral para caracterização de *outliers*, ilustrado na cor azul.

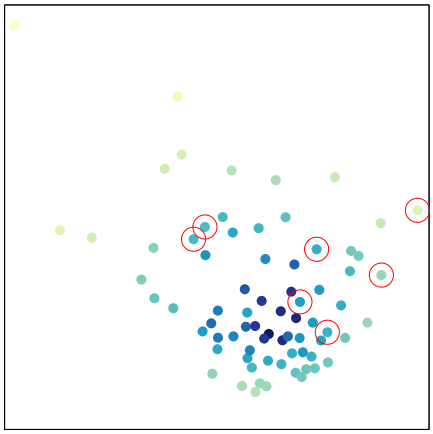


Fonte: Elaborada pelo autor.

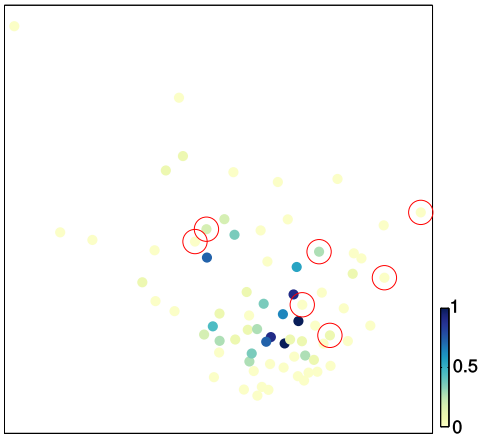
Desconhece-se por parte do autor estudos específicos na literatura voltados à medida de desvio de profundidade dos dados, fato também motivado pelo uso da medida de *stress* e da abordagem topológica do CheckViz nos experimentos. Embora pertencentes a diferentes classes de medidas de distorção, seu uso é motivado pela possibilidade de exploração de relações com a preservação de profundidade aqui proposta e discutida.

Além disso, visando o aperfeiçoamento dos resultados, o uso de outras técnicas de profundidade, como Random Tukey Depth (CUESTA-ALBERTOS; NIETO-REYES, 2008). Cuesta-Albertos e Nieto-Reyes (2008) argumenta que, devido a sua construção utilizar cálculos em subespaços, pode apresentar melhores resultados com o aumento da dimensão do dado. Essas diferenças são ilustradas na Figura 52, onde os *outliers* são de fato pontos de baixa profundidade, de acordo com essa função de profundidade, aparentando ser uma linha interessante de investigação.

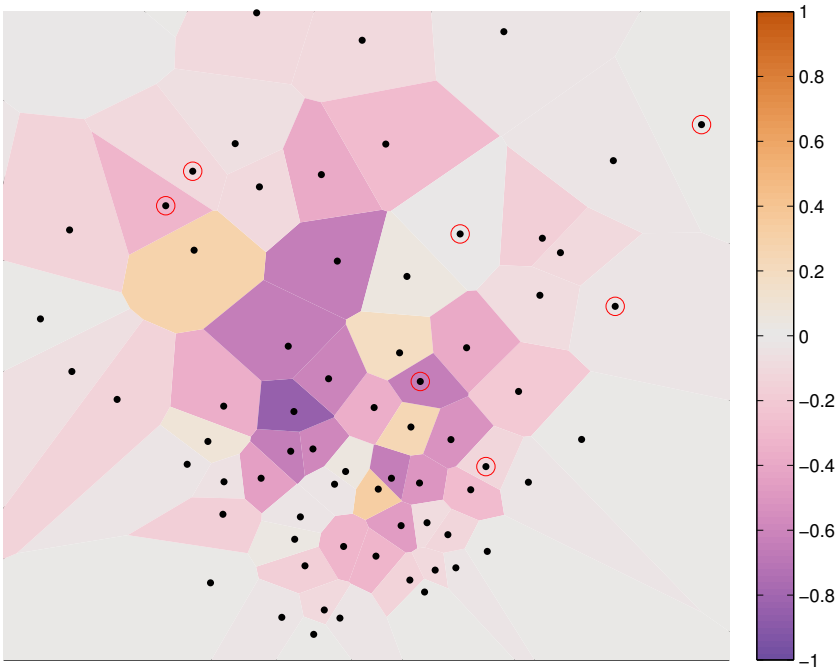
Figura 52 – Profundidade dos dados calculada no conjunto base de dados Hepatitis utilizando *Sammon mapping*.



(a) L_1D depth.



(b) *Random Tukey depth*.



(c) D^d avaliado utilizando *Random Tukey depth*.

Fonte: Elaborada pelo autor.

CAPÍTULO

4

CENTRALIDADE E TEORIA DE *KERNELS*

Teoria de *kernels* vem sendo utilizada em diversas áreas, incluindo extração de padrões (SCHöLKOPF *et al.*, 1999) e problemas de classificação em *Machine learning* (KULIS, 2013). Técnicas de projeção multidimensional que utilizam teoria de *kernels* também têm sido propostas, como a Análise de Componentes Principais com Kernel (KPCA) (SCHöLKOPF; SMOLA; MüLLER, 1998), KPCA com pontos de controle (c-KPCA) (OGLIC; PAURAT; GäRTNER, 2014) e mais recentemente a Kelp (BARBOSA *et al.*, 2016). Isso motiva analisar o comportamento de funções de centralidade que utilizem teoria de *kernels*, possibilitando seu posterior uso em combinação com tais técnicas de projeção.

Além disso, pode ser encontrado na literatura uma abordagem para detecção de novidades em problemas de classificação de uma classe (*one-class classification*), que relaciona-se com detecção de *outliers* e que utiliza-se do cálculo do Kernel PCA (HOFFMANN, 2007).

Assim, ao longo desse Capítulo investigamos uma noção de profundidade utilizando teoria de *kernels* e seu comportamento após a projeção.

4.1 Introdução

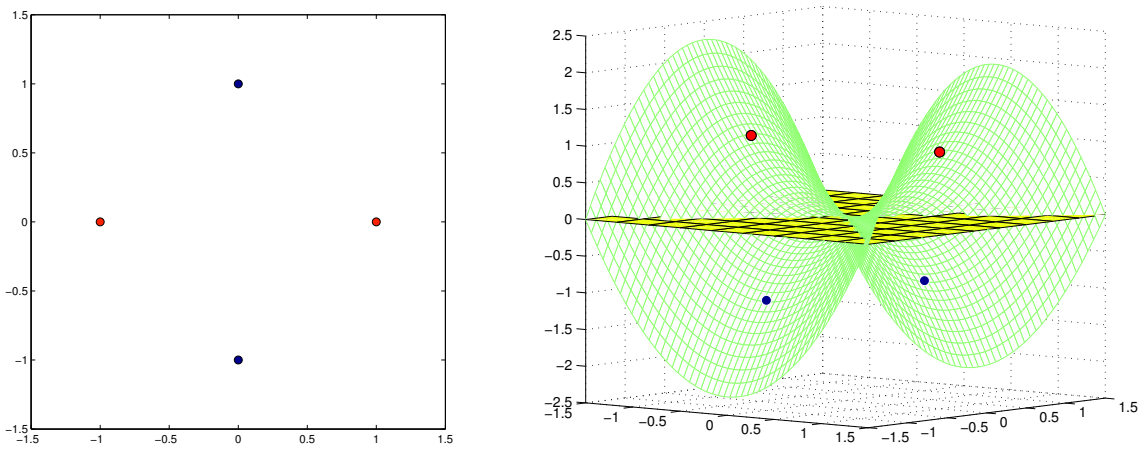
O uso de métodos de kernel é bastante comum na literatura de aprendizado de máquina, particularmente em aprendizado de métricas e classificação não-supervisionada (SCHöLKOPF; SMOLA; MüLLER, 1998), por sua possibilidade de lidar com não-linearidades entre os dados.

É sabido que a utilização de métodos lineares para solução desses problemas possui uma limitação intrínseca, ainda em casos simples, como ilustrado na Figura 53a. No caso ilustrado, nenhum classificador linear consegue separar perfeitamente os dados em regiões contendo os pontos de cada classe de forma disjunta, isto é, nenhuma reta separa apenas pontos vermelhos de um lado e apenas pontos azuis do outro.

Uma forma possível de resolver esse problema consistiria em efetuar um mapeamento

não-linear dos dados, para um espaço de dimensão maior, de forma que após o mapeamento um classificador linear seja suficiente para resolver o problema. No exemplo da Figura 53a, após efetuar o mapeamento não linear $z(x,y) = x^2 - y^2$ dos dados em $\mathbb{R}^2$, ilustrado na Figura 53b, o os dados passam a ser linearmente separáveis em $\mathbb{R}^3$ pelo plano $z = 0$, ilustrado pelo plano amarelo.

Figura 53 – Exemplo de limitação para classificadores lineares.



(a) (b)

Fonte: Elaborada pelo autor.

O controle de um mapeamento adequado pode ser uma escolha delicada por vários aspectos. Em primeiro lugar, a escolha de um mapeamento pode ser adequada para um tipo de dados e não adequada para outro, ou seja, diferentes fontes de não-linearidade nos dados. Em segundo lugar, a depender o mapeamento, a avaliação da métrica pode não ser viável computacionalmente, como será descrito em seguida.

Métodos de kernel possibilitam uma forma unificada, viável computacionalmente, flexível e elegante de lidar com tais problemas (SCHöLKOPF; SMOLA; MüLLER, 1998). Uma de suas principais vantagens consiste em fazer avaliações em tipos gerais de dados, por ex., dados vetoriais, *strings*, grafos, etc. Além disso, não possuem restrições quanto a dimensão inicial dos dados (HOFMANN; SCHÖLKOPF; SMOLA, 2008).

A intuição é análoga, ou seja, considerar que antes de medir as similaridades par-a-par entre dois elementos de um conjunto de dados, eles são mapeados para um espaço de maior dimensão - um espaço de Hilbert (*feature space*) - e a partir disso suas similaridades são medidas. Porém, a depender do mapeamento considerado, calcular medidas de similaridades nesse espaço de Hilbert $\mathscr{H}$ poderia ser inviável em termos computacionais, uma vez que sua dimensão não é necessariamente finita. Tomando a similaridade pelo produto interno, em tais espaços resultaria em avaliar uma integral em um espaço de dimensão infinita, por exemplo (SCHöLKOPF; SMOLA; MüLLER, 1998).

A ideia principal de métodos de kernel é contornar isso avaliando o produto interno em tais espaços de Hilbert, para onde os dados seriam mapeados, através de uma função κ

avaliada nos dados no espaço original, ou seja, antes do mapeamento. Mais formalmente, sejam $\{\mathbf{x_i}, \mathbf{x_j}\}$ dois pontos quaisquer do seu conjuntos de dados, $\{\mathbf{x} \in X \subset \mathbb{R}^m\}$, $\phi : X \to \phi(X) \subset \mathscr{H}$ um mapeamento tal que

$$\langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle_{\mathscr{H}} = \kappa(\mathbf{x_i}, \mathbf{x_j}) \ , \tag{4.1}$$

a função κ é chamada de *função kernel*. O teorema de Mercer garante que a Equação 4.1 é bem definida para qualquer função κ que satisfizer as condições de Mercer, isto é, ser simétrica, positiva e semi-definida (SCHöLKOPF; SMOLA; MüLLER, 1998). Uma matriz $\mathbf{K}$ proveniente da avaliação em um conjunto finito de pontos dessa função κ no $\mathbb{R}^m$ satisfaz as condições do teorema se for simétrica, positiva e semi-definida (SCHöLKOPF *et al.*, 1999).

Cabe notar que pela igualdade da Equação 4.1, o cálculo explícito do mapeamento ϕ não se faz necessário para calcular o produto interno em $\mathscr{H}$, sendo implicitamente definido pela escolha da função kernel κ. A essa construção é dado o nome de *kernel trick* ou substituição de kernel (*kernel substitution*) e torna o cálculo do produto interno viável computacionalmente (BISHOP, 2006). Algumas funções kernel comuns, que possuem forma analítica, são listadas a seguir:

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \left(\langle \mathbf{x_i}, \mathbf{x_j} \rangle + c\right)^d \qquad \text{(Polinomial)} \ , \tag{4.2}$$

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \tanh\left(a\langle \mathbf{x_i}, \mathbf{x_j} \rangle + b\right) \qquad \text{(Sigmoidal)} \tag{4.3}$$

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\tfrac{1}{\sigma}\|\mathbf{x_i} - \mathbf{x_j}\|^2\right) \qquad \text{(Laplace)} \ , \tag{4.4}$$

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\tfrac{1}{2\sigma^2}\|\mathbf{x_i} - \mathbf{x_j}\|^2\right) \qquad \text{(Gaussiano)} \ , \tag{4.5}$$

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = \prod_{k=1}^{m} B_{2p+1}\left(x_i^k - x_j^k\right) \qquad \text{(B-Spline)} \ , \tag{4.6}$$

com $a, b, c, \sigma \in \mathbb{R}$, $d, k, p, P \in \mathbb{N}$, $B_p := B_{p-1} \otimes B_0$, com

$$B_p(x) = \sum_{r=0}^{p+1} \frac{(-1)^p}{p!} \binom{p+1}{r} \left(x + \frac{p+1}{2} - r\right)^p ,$$

sendo x_i^k a k-ésima coordenada de $\mathbf{x_i} \in \mathbb{R}^m$ e $\otimes$ o operador de convolução (HOFMANN; SCHÖLKOPF; SMOLA, 2008).

Além dessas funções para dados vetoriais, no Quadro 2 são listadas funções *kernel* para outros tipos de dados.

A flexibilidade de tal estratégia vem da possibilidade de combinar funções kernel entre si definindo uma nova função kernel $\hat{\kappa}$. Dados dois pontos quaisquer $\mathbf{x_i}, \mathbf{x_j}$ do domínio e duas funções kernel $\kappa_1(\mathbf{x_i}, \mathbf{x_j}), \kappa_2(\mathbf{x_i}, \mathbf{x_j})$, algumas construções possíveis são:

Quadro 2 – Diferentes funções kernel para diversos tipos de dados.

Tipo de dados	Kernel	Referência
Strings	*Bag-of-Words* n-grams *Mismatch* Árvores de sufixo	(HOFMANN; SCHÖLKOPF; SMOLA, 2008) (HOFMANN; SCHÖLKOPF; SMOLA, 2008; COLLINS; DUFFY, 2001)
Conjuntos	*Subset kernel* ANOVA	(BISHOP, 2006)
Histogramas	Intersecção Generalizada *Pyramid match*	(BOUGHORBEL; TAREL; BOUJEMAA, 2005) (GRAUMAN; DARRELL, 2007)
Grafos	Kernel de difusão Laplaciano regularizado Entre grafos Caminho mais curto	(KONDOR; LAFFERTY, 2002) (SMOLA; KONDOR, 2003) (GÄRTNER, 2003; BORGWARDT *et al.*, 2005) (BORGWARDT; KRIEGEL, 2005)
Distribuição de prob.	Kullback-Leibler Chernoff Bhattacharyya Fisher	(ZHOU; CHELLAPPA, 2006) (BISHOP, 2006; HOFMANN; SCHÖLKOPF; SMOLA, 2008)

Fonte: Elaborada pelo autor.

$$\hat{\kappa}(\mathbf{x_i},\mathbf{x_j}) = c\kappa_1(\mathbf{x_i},\mathbf{x_j}) \ , \tag{4.7}$$

$$\hat{\kappa}(\mathbf{x_i},\mathbf{x_j}) = f(\mathbf{x_i})\kappa_1(\mathbf{x_i},\mathbf{x_j})f(\mathbf{x_j}) \ , \tag{4.8}$$

$$\hat{\kappa}(\mathbf{x_i},\mathbf{x_j}) = q(\kappa_1(\mathbf{x_i},\mathbf{x_j})) \ , \tag{4.9}$$

$$\hat{\kappa}(\mathbf{x_i},\mathbf{x_j}) = \kappa_1(\mathbf{x_i},\mathbf{x_j}) + \kappa_2(\mathbf{x_i},\mathbf{x_j}) \ , \tag{4.10}$$

$$\tag{4.11}$$

com $c \in \mathbb{R}^+, f(.)$ uma função qualquer, e $q(.)$ um polinômio com coeficientes não-negativos. Para mais formas de construção de novas funções *kernel* o leitor pode consultar o livro de Bishop (2006).

A miríade de possíveis construções acaba introduzindo uma dificuldade na escolha de qual função *kernel* utilizar, ou ainda sim de como ajustar seu eventuais parâmetros - chamados de hiperparâmetros. Uma outra forma de abordar o problema, além do escopo deste trabalho, consiste em aprender uma combinação de *kernels* que obedeça restrições de similaridade e dissimilaridade entre os dados, através de regras (*constraints*) definidas a priori. Essa construção é uma parte da área de aprendizagem de métricas, conhecida como aprendizagem de *kernels* (ABBASNEJAD; RAMACHANDRAM; MANDAVA, 2012).

A elegância vêm do fato que *todo* algoritmo linear que dependa apenas da medida de similaridade entre os pontos - como avaliações do produto interno - pode ter sua versão não-linear. Para isso, suas equações com produto interno devem ser reescritas considerando o mapeamento implícito ϕ, e utilizando o *kernel trick*, com o objetivo de ter as equações descritas apenas em termos da função kernel κ, o que dá a versão *kernelizada* do algoritmo linear.

A flexibilidade decorre do fato que para analisar diferentes comportamentos não-lineares induzidos pela ϕ não é necessário modificar o pipeline do algoritmo utilizado, mas apenas a

função kernel utilizada.

Para um *insight* geométrico sobre a não-linearidade introduzida Burges (1998) analisa a estrutura geométrica de $\mathscr{H}$ e mostra que a análise da métrica Riemanniana induzida por ϕ pode ser feita inteiramente em termos da função *kernel* associada. É feita apenas a análise do caso para funções *kernel* polinomiais (BURGES, 1998).

4.2 Kernel PCA

Uma das primeiros técnicas a ter sua versão kernelizada foi a Análise de Componentes Principais (PCA), definindo o *Kernel PCA*, utilizada como uma técnica de redução de dimensionalidade. Utiliza-se do mesmo princípio descrito anteriormente, que corresponderia a efetuar o mapeamento implícito dos dados para um espaço de Hilbert $\mathscr{H}$ e aplicar o PCA nos dados mapeados (SCHöLKOPF *et al.*, 1999). Similarmente ao caso linear, após a aplicação do *kernel trick*, o seguinte problema de auto-valor precisa ser resolvido

$$\mathbf{K}\alpha = l\lambda\alpha \ , \tag{4.12}$$

onde a matriz do kernel $\mathbf{K}$ é definida tal que $\mathbf{K}_{ij} = \kappa(\mathbf{x_i}, \mathbf{x_j})$, l é a cardinalidade do conjunto de dados, e o espectro da matriz $\mathbf{K}$ é modificado por um fator de escala de $\sqrt{\lambda_i}$. Os autovetores de $\mathbf{K}$ correspondem às componentes principais em $\mathscr{H}$.

Na Figura 54, é possível ilustrar um exemplo de projeção multidimensional aplicando o PCA e sua versão kernelizada KPCA, que será melhor detalhada em seguida. Vale notar também que, o PCA pode ser visto como o Kernel PCA utilizando a função kernel $\kappa(\mathbf{x_i}, \mathbf{x_j}) = \langle \mathbf{x_i}, \mathbf{x_j} \rangle$, a menos de um fator de escala.

Para ilustrar a *kernelização* de um algoritmo linear, vamos avaliar duas construções distintas do Kernel PCA.

Na primeira, seja um conjunto de n pontos $X = \{\mathbf{x}_i\} \subset \mathbb{R}^m$ define um mapeamento não linear ϕ na forma

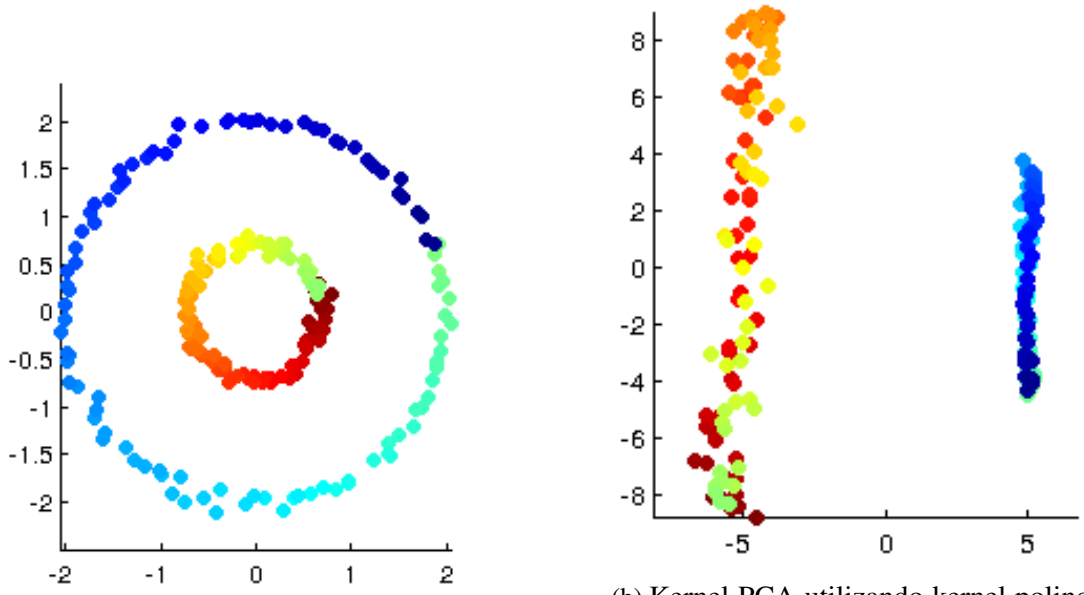
$$\begin{aligned} \phi : X \subset \mathbb{R}^m &\longrightarrow \mathscr{H} \\ \mathbf{x}_i &\longmapsto \phi(\mathbf{x}_i) = \phi_i \ , \end{aligned}$$

O *PCA* é efetuado nos pontos $\phi(X)$, ou seja, determina-se a matriz de covariância

$$\mathbf{C} = \frac{1}{n} \sum_{j=1}^{n} \phi_j \phi_j^T \ .$$

Considere $\mathbf{v} \in \mathscr{H}$ um auto-vetor de $\mathbf{C}$, e seja $\lambda > 0$ seu autovalor correspondente, $\mathbf{v} \in span\{\phi_1, \ldots \phi_n\}$, logo $\mathbf{v} = \sum_{i=1}^{n} \alpha_i \phi_i$. Assim, como pode ser visto em Schölkopf, Smola e

Figura 54 – Projeções utilizando PCA e Kernel PCA em um conjunto de dados com dois círculos concêntricos definidos em $\mathbb{R}^2$.



(a) PCA.

(b) Kernel PCA utilizando kernel polinomial de grau 2 (d=2).

Fonte: Elaborada pelo autor.

Müller (1998), para efetuar a decomposição espectral de $\mathbf{C}$, o mapeamento ϕ não precisa ser calculado explicitamente, pois

$$\lambda \mathbf{v} = \mathbf{C}\mathbf{v} \quad \Rightarrow \quad \lambda \langle \phi_k, \mathbf{v} \rangle = \langle \phi_k, \mathbf{C}\mathbf{V} \rangle \ , \tag{4.13}$$

$$\Rightarrow \quad \lambda \left\langle \phi_k, \sum_{i=1}^{n} \alpha_i \phi_i \right\rangle = \left\langle \phi_k, \frac{1}{n} \sum_{j=1}^{n} \phi_j \phi_j^T \mathbf{v} \right\rangle, \forall k = 1, \ldots, n \ , \tag{4.14}$$

$$\Rightarrow \quad \lambda \sum_{i=1}^{n} \langle \phi_k, \phi_i \rangle \alpha_i = \frac{1}{n} \left\langle \phi_k, \sum_{j=1}^{n} \langle \phi_j, \mathbf{v} \rangle \phi_j \right\rangle \forall k = 1, \ldots, n \ , \tag{4.15}$$

$$\Rightarrow \quad \mathbf{K}\alpha = \frac{1}{\lambda n} \mathbf{K}^2 \alpha \ , \tag{4.16}$$

$$\Rightarrow \quad (n\lambda)\alpha = \mathbf{K}\alpha \ , \tag{4.17}$$

se $\mathbf{K}$ tiver posto completo, com $\mathbf{K}_{ij} = \left\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \right\rangle = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \ldots, n$. Essa construção é particularmente interessante pois observe que podemos determinar a decomposição espectral de $\mathbf{C}$, resolvendo o problema de autovalor da Equação 4.17.

Para determinar as coordenadas da projeção de um ponto $\mathbf{z}$ qualquer em uma k-ésima componente principal, $k \leq n$, tem-se

$$B_k = \langle v^k, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{l} a_i^k \kappa(\mathbf{x_i}, \mathbf{x}) \ , \tag{4.18}$$

ou seja, utiliza-se apenas apenas as avaliações de κ nos pontos do espaço original $\mathbf{x}_i \in \mathbb{R}^m$ e os coeficientes de α, que são autovetores da matriz $\mathbf{K}$ - avaliação da função kernel nos pontos do conjunto.

Assim como o PCA, o Kernel PCA precisa ter os dados centralizados. Schölkopf, Smola e Müller (1998) mostram que para centrar os dados em $\mathscr{H}$, isto é $\tilde{\phi}(\mathbf{x_i}) = \phi(\mathbf{x_i}) - \bar{\phi}$, com $\bar{\phi}$ a média dos ϕ_i, basta substituir $\mathbf{K}$ por

$$\widetilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_n\mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n \ ,$$

onde $(\mathbf{1}_n)_{ij} = \frac{1}{n}$.

Para a segunda construção, à luz da otimização, dado um conjunto $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, para cada ponto $\mathbf{x}_i \in X \subset \mathbb{R}^m$, efetuar o PCA corresponde a determinar e ordenar direções $\mathbf{w}$ de acordo com as quais a variação das projeções $\mathbf{w^T x}$, decresça, ou seja, na primeira componente principal de $\mathbf{C}$ encontra-se a direção na qual as projeções mais variam.

Assim, para o conjunto de pontos X, passa-se a resolver o problema de otimização com restrições

$$\max_{\mathbf{w}} \frac{1}{n} \sum_{\mathbf{x}_j \in X} (\mathbf{w}^T \mathbf{x}_j)^2 = \max_{\mathbf{w}} \ \mathbf{w}^T \mathbf{C} \mathbf{w} \ ,$$
$$\text{tal que } \mathbf{w}^T \mathbf{w} = 1.$$

Ao considerar todas as direções principais, os autovetores de $\mathbf{C}$ definem uma nova base rotacionada de acordo com a variância dos dados e as projeção dos dados, suas coordenadas nessa nova base.

Ao descartar componentes (redução de dimensionalidade), o PCA pode ser utilizado como remoção de ruído, pois as projeções não consideram as componentes principais que possuem menor variação.Essa estratégia funciona bem quando os dados utilizados são *lineares*.

Ao considerar a análise de componentes principais com *kernel*, Suykens, Gestel e Brabanter (2002) mostram que ela é equivalente ao problema de otimização com restrições da Equação 4.19

$$\min_{\mathbf{w},\mathbf{e}} \frac{\gamma}{2n} \sum_{i=1}^{n} e_i^2 - \frac{1}{2}\mathbf{w}^T \mathbf{w} \ , \tag{4.19}$$
$$\text{tal que } e_i = \mathbf{w}^T(\phi(\mathbf{x_i}) - \hat{\mu}_\phi), \ i = 1, \cdots, n \ . \tag{4.20}$$

com $\gamma \in \mathbb{R}^+$ uma constante de regularização, $\hat{\mu_\phi} = (1/n)\sum_{j=1}^{n} \phi(\mathbf{x_j})$. O termo e_i^2 pode ser visto como uma função de custo quadrática $L(e_i) = e_i^2$, com isso, outras funções de custo podem ser abordadas.

Ao calcular as condições de otimalidade das equações (4.19) e (4.20) , sua solução corresponde a resolver um problema de autovalor na forma

$$\widetilde{\mathbf{K}}\alpha_{\mathbf{l}} = n\lambda_l \alpha_{\mathbf{l}}, \tag{4.21}$$

com $\widetilde{\mathbf{K}}$ sendo a matriz centrada do *kernel* e $\alpha_{\mathbf{l}}$ os coeficientes da restrição de igualdade do Lagrangiano, na equação (4.20). Dessa forma, pelo princípio da dualidade, resolver o problema

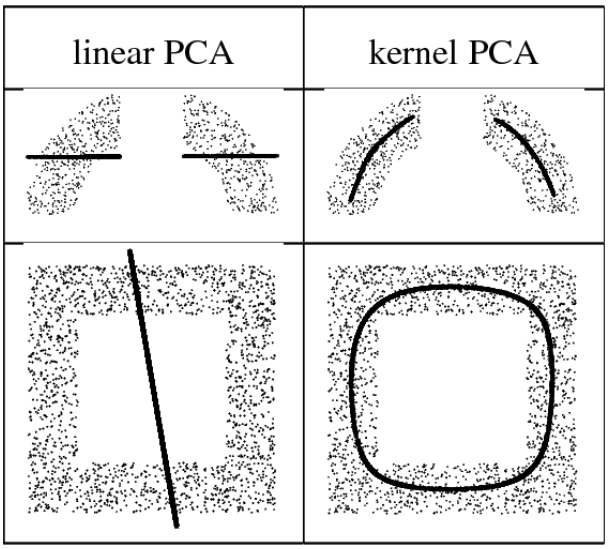
primal de otimização (4.19) é equivalente a resolver o problema dual (4.21) - em relação às variáveis introduzidas no Lagrangiano.

Suykens, Gestel e Brabanter (2002) mostram que o problema dual como um problema de autovetor na forma da equação (4.21), aparece apenas para uma função de custo L quadrática. Apesar de fora do escopo deste trabalho, o comportamento com outras funções de custo, eventualmente mais robustas, pode ser explorado através da solução do problema primal.

A matriz de Gram $\widetilde{\mathbf{K}}$ possui o mesmo espectro da matriz de covariância $\mathbf{C}$ de $\phi(X)$ em $\mathscr{H}$, a menos da normalização dos autovetores α_l por $1/\sqrt{\lambda_l}$, seus respectivos autovalores. Vale lembrar que, diferente do *PCA* onde a análise é feita linearmente para um subconjunto dos pontos, no *KPCA* essa análise, possibilita lidar com possíveis não linearidades nos dados. Na Figura 55 é ilustrada a diferença entre as componentes principais utilizando PCA e Kernel PCA.

A noção de subespaço definido pelas componentes principais acaba sendo, em uma primeira análise, contra-intuitiva no caso do Kernel PCA. Porém, é importante observar que o subespaço é definido em $\mathscr{H}$ e sua aproximação no espaço original ilustra a não-linearidade envolvida no processo. Cabe citar que a representação desse subespaço na Figura 55 é uma aproximação, uma vez que determinar *pré-imagens* de pontos em $\mathscr{H}$ após a projeção nas componentes principais não é imediata devido à diferença de dimensão entre o espaço original e $\mathscr{H}$ e à natureza implícita de ϕ. Apesar de diferentes aproximações na literatura, ainda é um problema em aberto (MIKA *et al.*, 1999; HONEINE; RICHARD, 2011). Além disso, esse problema difere de representar os pontos projetados como ilustrado na Figura 54b, que consiste apenas em obter as coordenadas $\{B_1, B_2\}$, descritos na Equação 4.18, para todos os pontos do conjunto de dados.

Figura 55 – Aproximação dos subespaços obtidos pela versão linear do PCA e sua versão com kernels.



Fonte: Adaptada de Schölkopf *et al.* (1999).

Observe que com o *PCA* tínhamos de resolver um problema de autovalor onde a matriz

tinha ordem m, dimensão do espaço de entrada. No *KPCA*, a matriz tem ordem da quantidade de pontos n, e possivelmente $n \gg m$.

Kernel PCA é um método de grande importância para diversas áreas, particularmente é um das primeiras variações de técnicas lineares (PCA) utilizando teoria de *kernels*. Zhang, Van Kaick e Dyer (2010) discutem como fazendo modificações na matriz pode-se explorar aplicações em processamento espectral de malhas. Adicionalmente, como pode ser visto na Seção 4.3, há uma ligação entre Kernel PCA e profundidade de dados utilizando *kernels*.

4.3 Funções de profundidade utilizando *Kernels*

O trabalho de Hu *et al.* (2011) propõe uma extensão de funções de profundidade de Mahalanobis por meio do uso da teoria de *kernels*. Inicialmente, uma análise do comportamento de funções de profundidade de Mahalanobis convencional é realizada considerando-se amostras esparsas. Para essas amostras, a matriz de covariância pode ser singular. Para o autor, o emprego desses estimadores padrão só faz-se justificável para casos em que as amostras geradas respeitam distribuições elípticas. Essa limitação motiva o desenvolvimento de uma função de profundidade denominada *Generalized Mahalanobis Depth* (GMHD) e sua extensão para uma versão com *kernels*, descrita como *kernel Generalized Mahalanobis Depth* (kmGMHD).

A função de profundidade que utiliza *kernels*, proposta por Hu *et al.* (2011), é definida como:

$$kmGMHD(\mathbf{x}) = \left[1 + \sum_{i=1}^{r} \frac{\left((\tilde{\kappa}(\mathbf{x},\mathbf{x}_j))^{j=1,\ldots,n}\alpha_i\right)^2}{\lambda_i^4}\right]^{-1}, \tag{4.22}$$

onde $\tilde{\kappa}(\mathbf{x},\mathbf{x}_j) = \kappa(\mathbf{x},\mathbf{x}_j) - \frac{1}{n}\sum_{l=1}^{n}\kappa(\mathbf{x},\mathbf{x}_l) - \frac{1}{n}\sum_{l=1}^{n}\kappa(\mathbf{x}_j,\mathbf{x}_l) + \frac{1}{n^2}\sum_{l=1}^{n}\kappa(\mathbf{x}_j,\mathbf{x}_l)$, que corresponde à centralização em $\mathscr{H}$; e $\{\lambda_i^2, \alpha_i\}$ são pares de autovalores não nulos e respectivos autovetores da matriz de *kernel* $\kappa(\mathbf{x}_k,\mathbf{x}_l)$.

Um aspecto interessante a ser observado é a possibilidade de levantamento de hipóteses sobre a distribuição modificando-se apenas a função de *kernel*, sem que exista a necessidade de modificação do *framework*. Nesse sentido, ao invés de se utilizar uma dada função *kernel* simples (por exemplo, gaussiana), pode-se definir uma função *kernel* em um espaço de curvas e então analisar toda a curva diretamente como um único elemento.

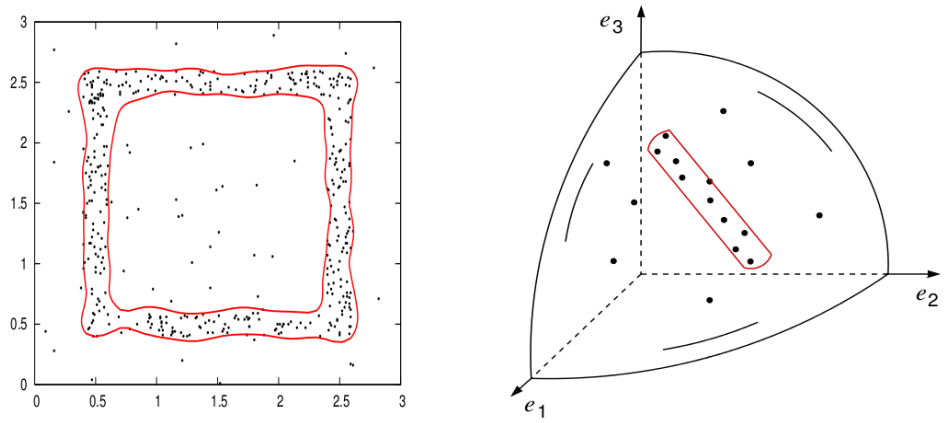
A construção segue a mesma intuição já descrita, assume que o dado $\mathbf{x}_i \in \mathbb{R}^m$ é implicitamente mapeado a um espaço de Hibert $\mathscr{H}$. A profundidade é então calculada nesse espaço, avaliando a função de kernel nos dados de entrada utilizando o *kernel trick*. Ele permite o cálculo da profundidade em $\mathscr{H}$ em função da escolha de uma *kernel function* $k(.,.)$. A intuição associada ao procedimento permite a identificação de possíveis não linearidades nos dados geradas pela influência da função de *kernel*.

Embora procedimentos bem definidos para a construção de novas funções *kernel* partindo-

se de antigas existam, como alguns descritos na Seção 4.2, tornando ampla a janela de possíveis combinações e análises, Hu *et al.* (2011) argumentam que não é imediato assumir que essas novas funções definirão novas funções de profundidade. Considerando-se esse aspecto, a presente análise limita-se para funções kernel polinomial e gaussiana, também analisadas por Hu *et al.* (2011). Ambas capturam relativamente bem não linearidades presentes nos dados sem necessidade de um grande esforço na escolha de parâmetros, como pode ser observado na Figura 58.

Uma medida baseada em *kernels* empregada para detecção de novidades está intimamente relacionada com a distância dos dados e o respectivo plano PCA em $\mathscr{H}$ (HOFFMANN, 2007), ilustrada na Figura 56. Quando aplicada junto a funções kernel gaussianas, essa métrica é conhecida por apresentar uma relação muito próxima a estimativas janeladas de Parzen (*Parzen-window estimates*), estimador de densidade bastante utilizado na literatura.

Figura 56 – Em vermelho, pontos que possuem a mesma distância ao plano do Kernel PCA. À esquerda sua representação no espaço original e à direita no espaço de Hilbert associado ao Kernel PCA.



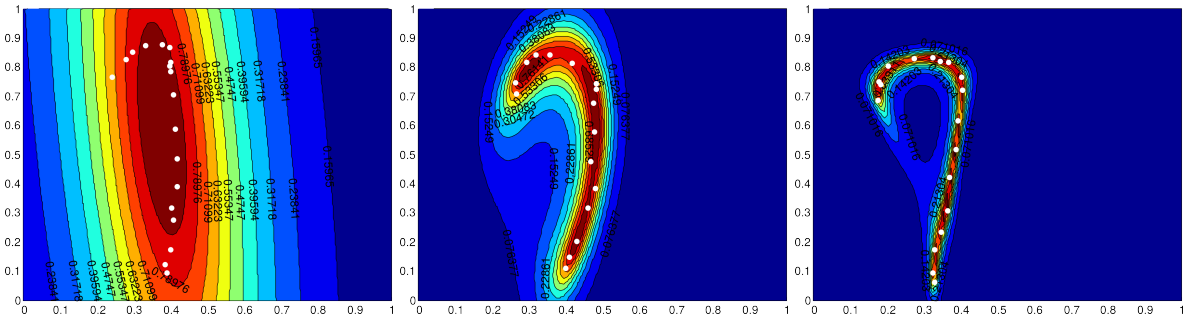
Fonte: Adaptada de Hoffmann (2007).

Ao analisar a Equação 4.22, a relação com outras técnicas fica aparente. Em particular, com Kernel PCA através do espectro da matriz de kernel. Porém, diferente do erro de reconstrução utilizado por Hoffmann (2007), não mede a distância em $\mathscr{H}$ para as componentes principais. De acordo com a generalização da profundidade de Mahalanobis proposta por Hu *et al.* (2011), a intuição seria a medida inversa da distância de Mahalanobis em $\mathscr{H}$ para o ponto mais central dos dados, isto é, $\bar{\phi}$.

Na Figura 57, cada isolinha representa um valor da *depth function*. Grandes valores da função indicam maior proximidade da região central.

Para função kernel polinomial foi escolhido $d = 2$. Notamos que alterando esse parâmetro não afetou significativamente nenhum dos resultados discutidos nessa tese. Para o kernel gaussiano, foi utilizado a heurística de Silvermann (SILVERMAN, 1986) para a escolha do σ,

Figura 57 – Conjunto de dados representados pontos brancos. Isovalores são apresentados para três funções de profundidade de dados diferentes.



(a) MHD padrão. (b) kmGMHD polinomial grau 2. (c) kmGMHD gaussiano.

Fonte: Elaborada pelo autor.

comumente usada em *Kernel Density Estimation - KDE*, descrita pela Equação 4.23

$$\sigma = 1.06\hat{\sigma}^{-1/5}, \tag{4.23}$$

onde $\hat{\sigma}$ é o desvio padrão médio das distâncias dos pontos.

Entretanto, nos testes efetuados só apresentou resultados razoáveis para dados de dimensão dois e três. Exemplos bidimensionais podem ser visto nas Figuras 57, 58 e 59.

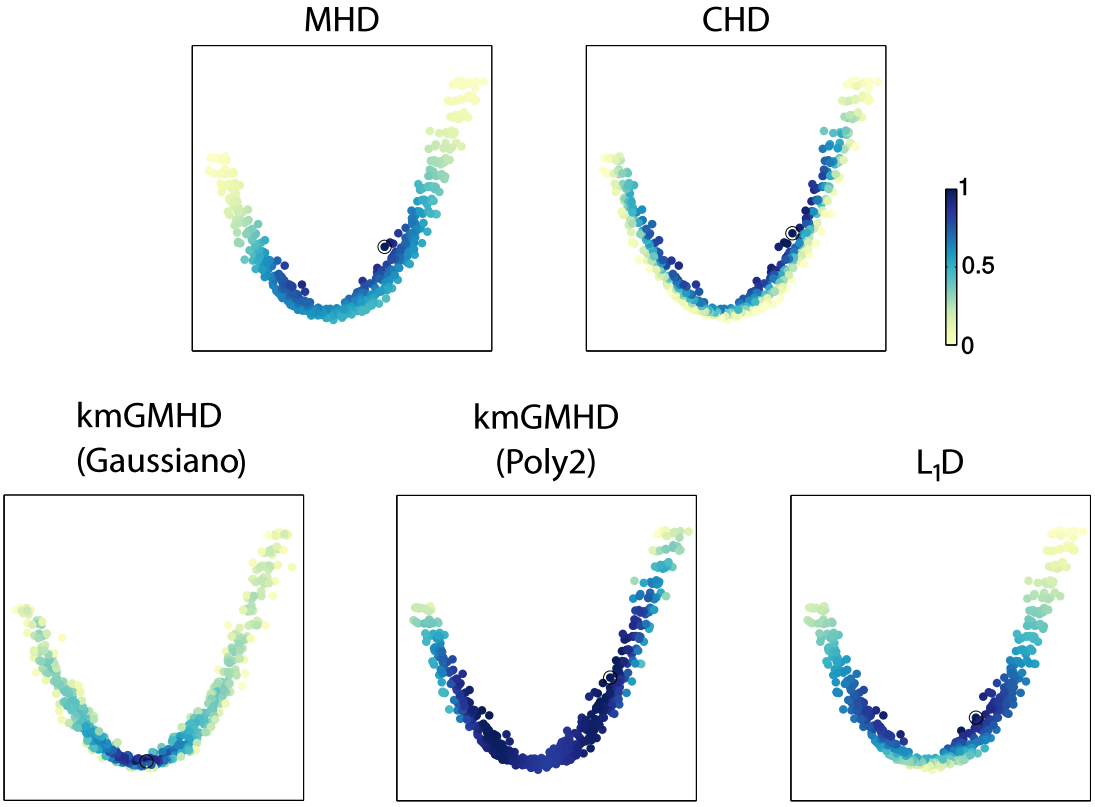
Na Figura 58 a função de profundidade utilizando *kernels* é comparada com as outras três citadas no Capítulo anterior em um conjunto de dados com uma estrutura não-linear. O circulo preto indica o ponto de maior profundidade de acordo com a função de profundidade utilizada.

Os mesmos testes envolvendo análise de robustez perante *outliers* foi efetuado nesse contexto, com os resultados ilustrados na Figura 59, entretanto utilizando *kmGMHD*. A possibilidade de alterar o comportamento da noção de centralidade através da modificação da função de kernel introduz uma flexibilidade na abordagem, porém introduzindo o custo de investigar diferentes parâmetros. Apesar de perder a característica não-paramétrica de outras funções de profundidade, efetuamos alguns experimentos que são discutidos ao longo desta Seção.

Como pode ser visto na Figura 59 a mudança da função kernel pode impactar significativamente no resultado. No caso do kernel gaussiano, a escolha do σ foi de acordo com a heurística de Silverman, produzindo resultados bem aceitáveis. Entretanto, acaba atribuindo valores intermediários de centralidade para a maioria dos outliers, o que destoa da robustez esperada.

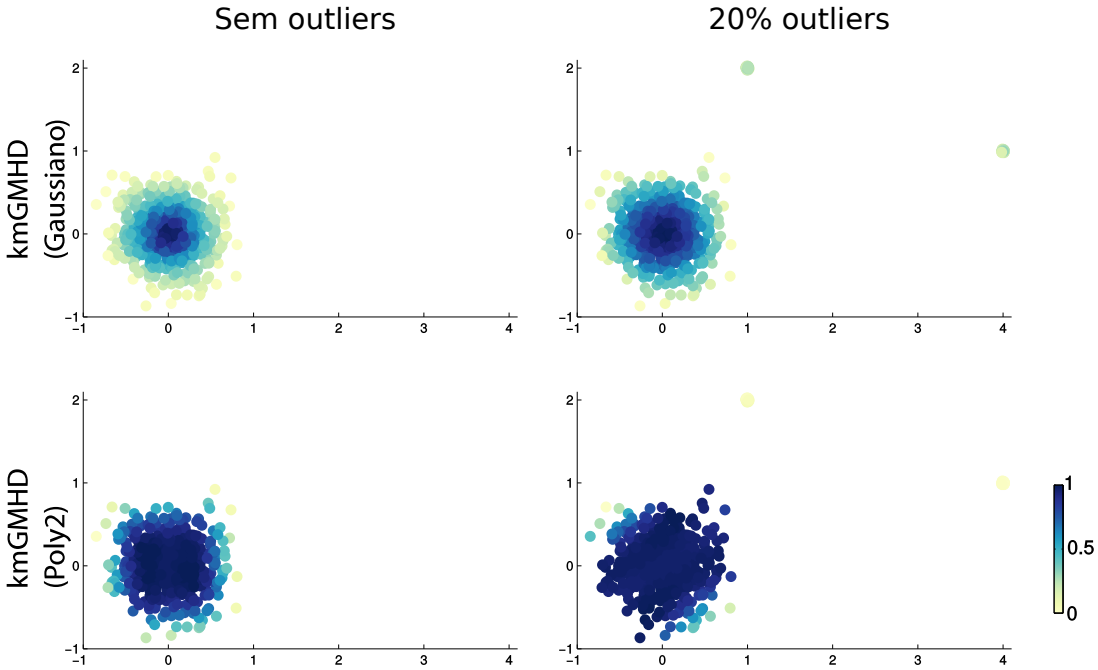
Note que por utilizar *kernels*, kmGMHD é escalável quanto a dimensão dos dados, uma vez que sua complexidade é alta no número de pontos a serem considerados, tornando-a atrativa para casos onde $m \gg n$, como descrito no Quadro 3. Entretanto, a escolha de kernel e parâmetros pode se tornar delicada em casos de dados multidimensionais mais gerais, etapa que requer uma

Figura 58 – Profundidade calculada em um conjunto de dados em forma com distribuição não-linear em forma de parábola. O círculo preto ilustra ponto de maior centralidade.



Fonte: Elaborada pelo autor.

Figura 59 – Avaliação de robustez na presença de outliers, colorido através da profundidade calculada utilizando kernels.



Fonte: Elaborada pelo autor.

investigação mais a fundo, tendo a heurística de Silverman como um ponto de partida para a análise.

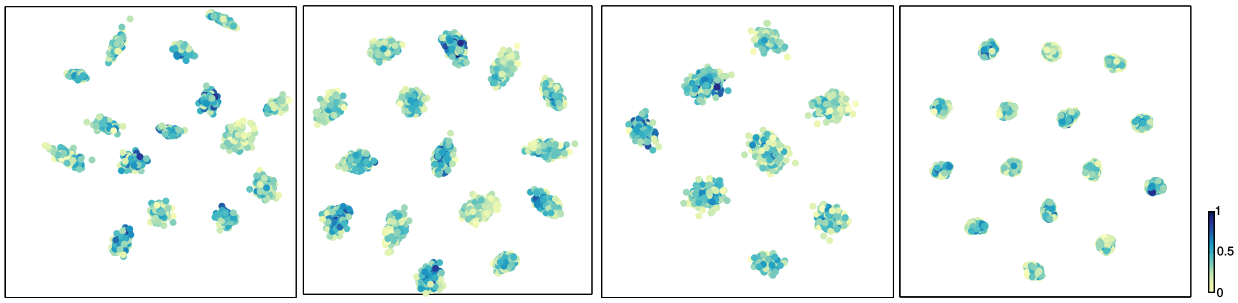
Quadro 3 – Algumas características das funções de profundidade analisadas.

Data depth	**Robustez a outliers**	**Complexidade assintótica**
L_1D	Sim	$O(n^2+nm)$
kmGMHD	Não	$O(n^3+m)$

Fonte: Elaborada pelo autor.

Na Figura 60 são ilustrados os valores de profundidade no espaço original (D^n) para o conjunto de dados *Ad10*, com *kernel* polinomial de grau 2. Já na Figura 61, é ilustrado o comportamento de D^2, que acaba não sendo muito intuitivo de analisar, no caso bidimensional.

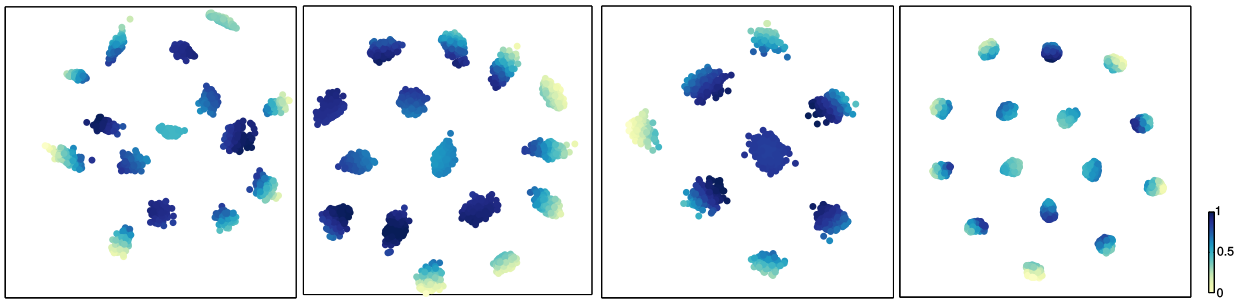
Figura 60 – D^m utilizando kmGMHD com kernel polinomial de grau 2 e conjunto de dados Ad10.



(a) PCA. (b) Sammon. (c) ICA. (d) t-SNE.

Fonte: Elaborada pelo autor.

Figura 61 – D^2 utilizando kmGMHD com kernel polinomial de grau 2 e conjunto de dados Ad10.



(a) PCA. (b) Sammon. (c) ICA. (d) t-SNE.

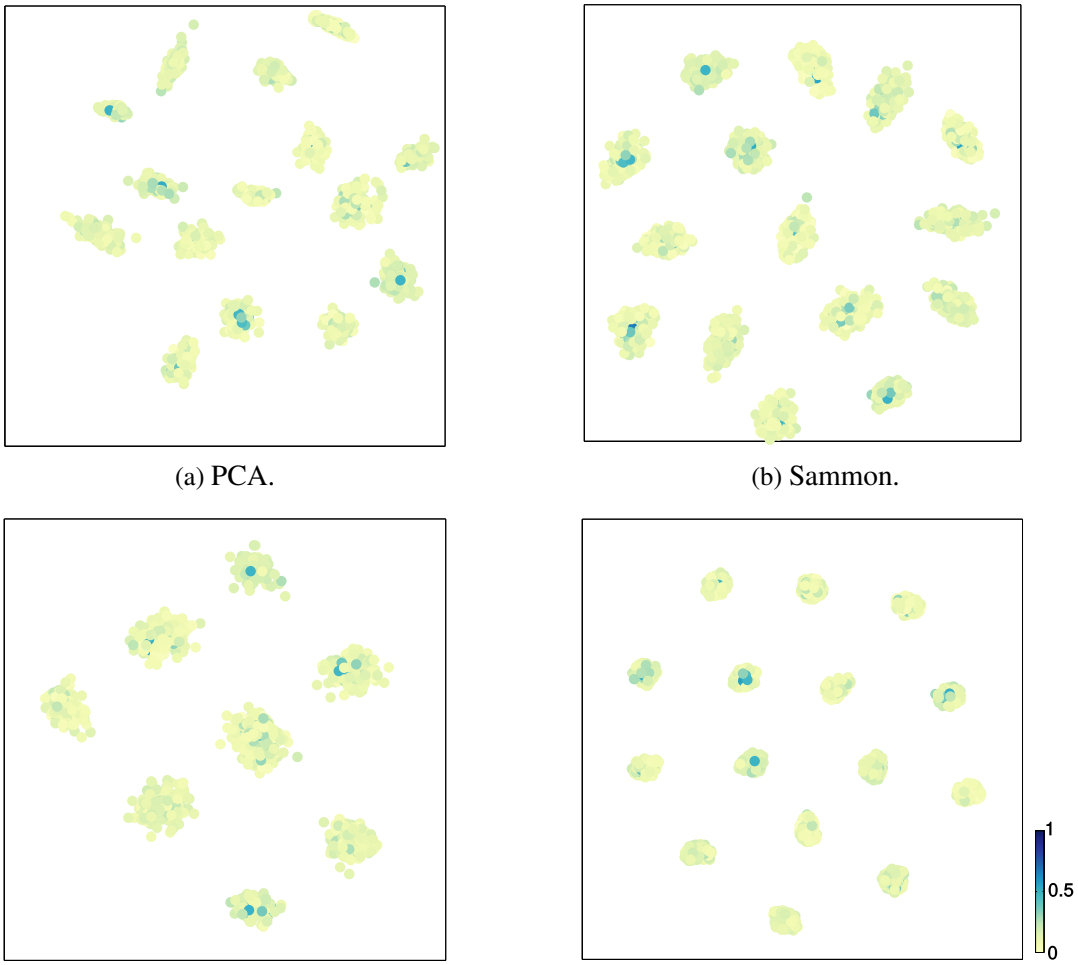
Fonte: Elaborada pelo autor.

Ao investigar esse mesmo conjunto de dados utilizando o kernel gaussiano com heurística de Silverman, obtemos o resultado ilustrado na Figura 62, que apresentam um resultado interessante, uma vez que em cada cluster há um ponto de centralidade mais alta e a distribuição de profundidade aparentar ser intra-clusters. Esse fato fica mais evidente no cálculo de D^2, ilustrado na Figura 63. Os valores de preservação de profundidade estão listados na Tabela 5, que

ilustram um outro comportamento: t-SNE como sendo a técnica que mais preserva para a noção de profundidade utilizando *kernels*. É possível que isso decorra da característica da técnica em preservar clusters, mas algo ainda para ser explorado.

De forma adicional, para técnicas de projeção multidimensional baseadas em funções *kernel* (OGLIC; PAURAT; GÄRTNER, 2014; BARBOSA *et al.*, 2016), a única estimativa de profundidade disponível para esse tipo de cenário é a kmGMHD.

Figura 62 – D^m utilizando kmGMHD com kernel gaussiano e conjunto de dados Ad10.

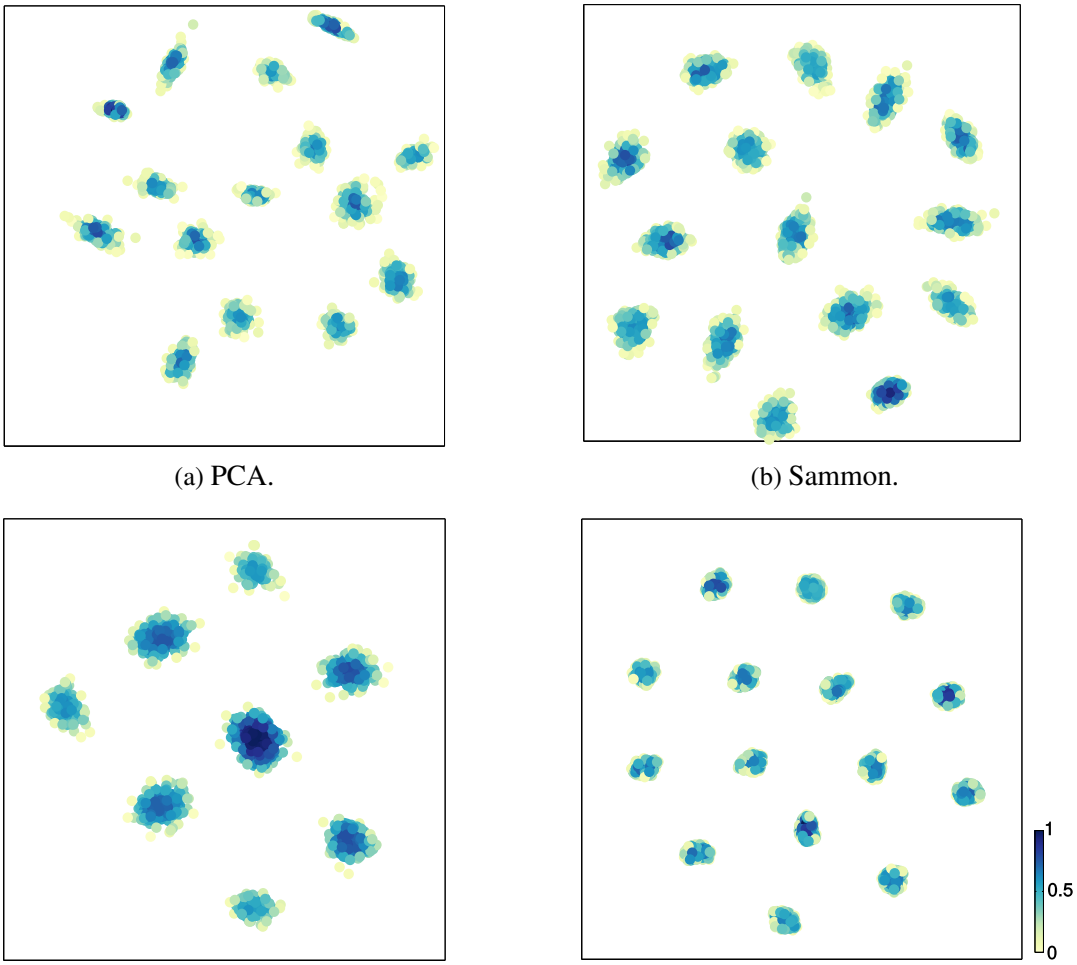


(a) PCA. (b) Sammon.

(c) ICA. (d) t-SNE.

Fonte: Elaborada pelo autor.

Figura 63 – D^2 utilizando kmGMHD com kernel gaussiano e conjunto de dados Ad10.



(a) PCA.

(b) Sammon.

(c) ICA.

(d) t-SNE.

Fonte: Elaborada pelo autor.

Tabela 5 – Análise quantitativa entre diferentes técnicas de projeção multidimensionais (melhores resultados são apresentados em negrito).

Dataset	**n**	**m**	**PCA**	**Sammon**	**ICA**	**t-SNE**
Ad10 - Gauss	1499	10	15.51	**13.89**	20.80	15.79
Ad10 - Poly2	1499	10	14.63	16.18	17.52	**11.71**
Faces - Gauss	697	4096	8.46	7.83	15.85	**7.44**
Faces - Poly2	697	4096	19.01	19.88	24.13	**17.59**

Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÃO

5.1 Discussão

Este trabalho analisou inicialmente uma nova metáfora visual - MIST - para visualização de coleções de documentos textuais, que possibilita a análise simultânea em vários níveis de informação em um único *layout*. O método utiliza simulação de corpo rígido para posicionar documentos no espaço visual minimizando a sobreposição dos elementos, enquanto busca preservar estruturas de vizinhança presentes no espaço original dos dados. A construção proposta possibilita a exploração em multi-nível de elementos da coleção que individualmente não possuem um elevado grau de relevância (centralidade), calculada sobre o grafo de co-autoria.

Foram realizadas comparações visuais e quantitativas que mostram que a MIST produz *layouts* mais compactos e preserva melhor a relação de vizinhança do que outras técnicas puramente baseadas em força. O layout é integrado com nuvens de palavras com objetivo de sumarizar os conjuntos de dados da coleção de forma a guiar o usuário para possíveis grupos de interesse para serem explorados de maneira interativa.

Além disso, a estratégia de considerando na simulação física apenas dados visíveis possibilita explorar a escalabilidade no processo, possibilitando uma exploração mais flexível de coleções maiores de documentos. Um outro aspecto que pode ser avaliado é como a adição incremental de pontos alteraria os resultados da MIST.

Ao longo da tese três aspectos de centralidade foram utilizados: grafos direcionados de citação (PageRank), estatistica não-paramétrica e centralidade utilizando teoria de *kernels*. A relação de citação entre documentos científicos constitui um grafo naturalmente. Porém, em casos onde isso não acontece, outras medidas de centralidade - como utilizando funções de profundidade definidas em grafos - pode ser um caminho de investigação (HUGG *et al.*, 2006). Assim, a MIST pode ser modificada para determinar a relevância de acordo.

Uma vez que o mapeamento de dados multidimensionais para o espaço visual, por meio

do uso de uma técnica de projeção, implica em perda de informação relativa aos dados originais, medidas de qualidade objetivam auxiliar na análise de eventuais distorções introduzidas no processo. Assim, o cálculo de profundidade de dados nos espaços original e visual permite uma forma de avaliação dessa técnica projeção a luz da medida de preservação de centralidade. Como pôde ser visto no Capítulo 3, a utilização dessa abordagem possibilita um novo entendimento ao analisar gráficos de dispersão com dados resultantes de projeção multidimensional.

Essa investigação é de grande importância por dois motivos: sua relação com a estimativa de *outliers* de valor extremo; e, como pontuado por Tatu (2013), a investigação sobre a relação entre a percepção do usuário e a métricas de qualidade de projeção ainda constitui um campo pouco explorado. Nesse contexto, estudos com usuário podem ser realizados para se avaliar como a centralidade medida através da profundidade dos dados relaciona-se com a percepção humana.

Nesse contexto, as um dos objetivos propostos nessa tese foi abordado pelos questionamentos Q1 e Q2, do Capítulo 3, onde foi estudado uma forma quantitativa e qualitativa de utilização de profundidade de dados para analisar projeções multidimensionais.

Adicionalmente, no contexto do questionamento Q3, três diferentes esquemas para a seleção de pontos de controle são propostos, utilizando o aspecto numérico da profundidade de dados e o aspecto semântico associado à centralidade e *outliers*.

Definir campos escalares sobre os dados traz benefícios, porque permite ampliar as formas possíveis de análise da modificação dos dados no processo de projeção. Por exemplo, permite efetuar a análise exploratória nos dados através da avaliação pela distribuição do campo escalar pelas dimensões do dado - utilizando o canal de cor, como profundidade, em coordenadas paralelas. Além disso, outras estratégias para medir a variação desse campo escalar, como persistência topológica, podem ser usadas também (RIECK; LEITTE, 2015).

No aspecto de utilização de teoria de *kernels* e sua relação com centralidade de dados, alguns problemas podem ser apontados, tais como: estimação de hiper-parâmetros, eventual não-robustez para uma função *kernel* escolhida. Porém, a flexibilidade de aprendizagem de novos *kernels* pode abrir uma análise interessante, tal que reflitam profundidade de dados como percebidas por analistas em um gráfico de dispersão. Possibilitam também analisar, por ex., a centralidade em textos através da utilização de *kernels* para *strings*, abrindo novas análises para coleções de documentos, por exemplo. Além disso, possibilita a utilização dessa medida de qualidade para técnicas de projeção que utilizam teoria de *kernels*.

5.2 Direções de investigação

No aspecto de posicionamento de discos no espaço visual, outros esquemas de otimização podem ser explorados, como o proposto por Gomez-Nieto *et al.* (2016). Uma vez que pode

possibilitar bons resultados tanto para elementos de tamanhos diferentes para diferentes direções (*word clouds*), sem o efeito de *stacking*, quanto para os elementos de tamanhos iguais em todas direções (discos), preservando compacidade e *layouts* originais.

Yma das grandes vantagens de fazer essa conexão previamente não explorada entre duas áreas, funções de profundidade de dados e projeção de dados multidimensionais, consiste na miríade de possibilidades que surgem para análise. Um importante aspecto que pode ser futuramente explorado refere-se à análise de como a profundidade dos dados é distribuída em cada dimensão dos dados, no espaço original. O uso de metáforas de janelas combinadas (*linked-views*) de gráficos de dispersão e coordenadas paralelas poderia definir um bom ponto de partida para essas análises.

A seguir listaremos mais algumas direções a serem futuramente investigadas.

Técnicas de projeção multidimensional

- Detecção de pontos de controle relevantes utilizando profundidade combinada com funções *hash* (SHAPIRA; AVIDAN; SHAMIR, 2009) e comparação de como isso afeta o layout da projeção;
- Uma técnica de projeção multidimensional cujo objetivo seja minimizar a distorção de um campo escalar definido sobre os pontos, por ex., a sua profundidade;
- Analisar o comportamento de profundidade em conjuntos de dados esparsos (por ex., documentos de texto);
- Modificar a centralidade calculada no grafo de citações utilizada na MIST considerando medidas de profundidade de dados para grafos (HUGG *et al.*, 2006);
- Avaliar a modificação da etapa de remoção de sobreposição pela proposta recentemente por Gomez-Nieto *et al.* (2016);
- Investigar a análise de centralidade por *clusters* utilizando *kernel* gaussiano e a função de profundidade com *kernel* (kmGMHD) para definição de pontos de controle, com o intuito de guiar projeções em técnicas semi-supervisionadas.

Profundidade de dados e Percepção

- Estimar centralidade em gráficos de dispersão segundo a taxonomia proposta por Sedlmair *et al.* (2012);
- Investigar as diferenças em casos bidimensionais entre a profundidade calculada e percepção humana sobre centralidade, seja via ferramentas interativas como *brushing* ou acompanhamento visual (*eye-tracking*).

Análise de Coleções Geométricas

- Explorar a noção de profundidade em coleções de entidades geométricas (curvas, superfícies, *patches* de superfícies) para a investigação de possíveis relações semânticas com partes de objetos.

REFERÊNCIAS

ABBASNEJAD, M. E.; RAMACHANDRAM, D.; MANDAVA, R. A survey of the state of the art in learning the kernels. **Knowledge and Information Systems**, Springer, v. 31, n. 2, p. 193–221, 2012. Citado na página 110.

AGGARWAL, C. **Outlier Analysis**. Springer New York, 2013. ISBN 9781461463962. Disponível em: <https://books.google.com.br/books?id=QQtGAAAAQBAJ>. Citado na página 76.

ALENCAR, A. B.; OLIVEIRA, M. C. F. de; PAULOVICH, F. V. Seeing beyond reading: a survey on visual text analytics. **WIREs Data Mining Knowl. Discov.**, v. 2, p. 476–492, 2012. Citado na página 43.

ALSAKRAN, J.; CHEN, Y.; LUO, D.; ZHAO, Y.; YANG, J.; DOU, W.; LIU, S. Real-Time Visualization of Streaming Text with a Force-Based Dynamic System. **IEEE Comput. Graph. Appl.**, v. 32, n. 1, p. 34–45, 2012. Citado 2 vezes nas páginas 46 e 47.

ANDREWS, K.; KIENREICH, W.; SABOL, V.; BECKER, J.; DROSCHL, G.; KAPPE, F.; GRANITZER, M.; AUER, P.; TOCHTERMANN, K. The InfoSky visual explorer: exploiting hierarchical structure and document similarities. **Information Visualization**, v. 1, n. 3/4, p. 166–181, 2002. Citado na página 45.

ARTHUR, D.; VASSILVITSKII, S. k-means++: The advantages of careful seeding. In: **Symposium on Discrete Algorithms**. [S.l.: s.n.], 2007. p. 1027–1035. Citado na página 58.

AUPETIT, M. Visualizing distortions and recovering topology in continuous projection techniques. **Neurocomputing**, v. 70, p. 1304–1330, 2007. ISSN 09252312. Citado 6 vezes nas páginas 14, 73, 74, 76, 85 e 90.

BARBOSA, A.; PAULOVICH, F.; PAIVA, A.; GOLDENSTEIN, S.; PETRONETTO, F.; NONATO, L. Visualizing and interacting with kernelized data. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 3, p. 1314–1325, 2016. Citado 2 vezes nas páginas 107 e 120.

BERGER, W.; PIRINGER, H.; FILZMOSER, P.; GROELLER, M. E. Uncertainty-aware exploration of continnuous parameter spaces using multivariate prediction. **Computer Graphics Forum**, v. 30(3), n. 3, p. 911–920, 2011. Citado 2 vezes nas páginas 27 e 32.

BERTINI, E.; TATU, A.; KEIM, D. Quality metrics in high-dimensional data visualization: An overview and systematization. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2203–2212, 2011. ISSN 10772626. Citado na página 35.

BEYER, K.; GOLDSTEIN, J.; RAMAKRISHNAN, R.; SHAFT, U. When is nearest neighbor meaningful? In: **International Conference in Database Theory**. [S.l.]: Springer, 1999. p. 217–235. Citado 2 vezes nas páginas 28 e 75.

BISHOP, C. **Pattern Recognition and Machine Learning**. Springer, 2006. (Information Science and Statistics). ISBN 9780387310732. Disponível em: <https://books.google.com.br/books?id=kTNoQgAACAAJ>. Citado 2 vezes nas páginas 109 e 110.

BORGWARDT, K. M.; KRIEGEL, H.-P. Shortest-path kernels on graphs. In: IEEE. **IEEE International Conference on Data Mining**. [S.l.], 2005. p. 8–pp. Citado na página 110.

BORGWARDT, K. M.; ONG, C. S.; SCHÖNAUER, S.; VISHWANATHAN, S.; SMOLA, A. J.; KRIEGEL, H.-P. Protein function prediction via graph kernels. **Bioinformatics**, Oxford Univ Press, v. 21, n. suppl 1, p. i47–i56, 2005. Citado na página 110.

BOSTOCK, M. D3. js. **Data Driven Documents**, 2012. Citado 2 vezes nas páginas 30 e 31.

BOUGHORBEL, S.; TAREL, J.-P.; BOUJEMAA, N. Generalized histogram intersection kernel for image recognition. In: IEEE. **IEEE International Conference on Image Processing**. [S.l.], 2005. v. 3, p. III–161. Citado na página 110.

BRINKE, W. ten; SQUIRE, D. M.; BIGELOW, J. Similarity: measurement, ordering and betweenness. In: SPRINGER. **International Conference on Knowledge-Based and Intelligent Information and Engineering Systems**. [S.l.], 2004. p. 996–1002. Citado na página 33.

BROWN, E. T.; LIU, J.; BRODLEY, C. E.; CHANG, R. Dis-function: Learning distance functions interactively. In: **IEEE Conference on Visual Analytics Science and Technology (VAST @ IEEE VIS)**. [S.l.]: IEEE, 2012. p. 83–92. ISBN 978-1-4673-4753-2. Citado na página 31.

BURGES, C. J. C. Geometry and lnvariance in Kernel Based Methods. In: SCHOLKOPF, B.; BURGES, C. J. C.; SMOLA, A. J. (Ed.). **Advances in Kernel Methods - Support Vector Learning**. [S.l.]: MIT Press, 1998. Citado na página 111.

CAMPOS, G. O.; ZIMEK, A.; SANDER, J.; CAMPELLO, R. J. G. B.; MICENKOVÁ, B.; SCHUBERT, E.; ASSENT, I.; HOULE, M. E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. **Data Mining and Knowledge Discovery**, p. 1–37, 2016. Citado 2 vezes nas páginas 87 e 88.

CAO, N.; SUN, J.; LIN, Y.-R.; GOTZ, D.; LIU, S.; QU, H. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 1172–1181, 2010. Citado na página 46.

CATTO, E. Iterative Dynamics with Temporal Coherence. **Game Developer Conference**, p. 1–24, 2005. Citado 2 vezes nas páginas 56 e 62.

CHEN, C. Top 10 unsolved information visualization problems. **IEEE Computer Graphics and Applications**, v. 25, n. 4, p. 12–16, 2005. ISSN 02721716. Citado na página 73.

CHUANG, J.; RAMAGE, D.; MANNING, C.; HEER, J. Interpretation and trust: designing model-driven visualizations for text analysis. In: **ACM Conference on Human Factors in Computing Systems (SIGCHI)**. [S.l.: s.n.], 2012. p. 443–452. Citado na página 48.

COLLINS, C.; CARPENDALE, S.; PENN, G. DocuBurst: Visualizing Document Content using Language Structure. **Computer Graphics Forum**, v. 28, p. 1039–1046, 2009. Citado na página 44.

COLLINS, C.; VIEGAS, F. B.; WATTENBERG, M. Parallel Tag Clouds to explore and analyze faceted text corpora. In: **IEEE Conference on Visual Analytics Science and Technology (VAST @ IEEE VIS)**. [S.l.: s.n.], 2009. p. 91–98. Citado na página 43.

COLLINS, M.; DUFFY, N. Convolution kernels for natural language. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2001. p. 625–632. Citado na página 110.

COUMANS, E. **Bullet 2.80 Physics SDK Manual**. [S.l.], 2012. Citado na página 61.

CUESTA-ALBERTOS, J. a.; NIETO-REYES, a. The random Tukey depth. **Computational Statistics and Data Analysis**, v. 52, n. 11, p. 4979–4988, 2008. ISSN 01679473. Citado na página 105.

CUI, W.; LIU, S.; TAN, L.; SHI, C.; SONG, Y.; GAO, Z.; QU, H.; TONG, X. TextFlow: Towards Better Understanding of Evolving Topics in Text. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2412–2421, 2011. Citado na página 44.

CUI, W.; WU, Y.; LIU, S.; WEI, F.; ZHOU, M.; QU, H. Context-Preserving, Dynamic Word Cloud Visualization. **IEEE Computer Graphics and Applications**, v. 30, p. 42–53, 2010. Citado na página 43.

DING, Y.; DANG, X.; PENG, H.; WILKINS, D. Robust clustering in high dimensional data using statistical depths. **BMC Bioinformatics**, v. 8 Suppl 7, p. S8, 2007. ISSN 14712105. Citado 2 vezes nas páginas 78 e 81.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *et al.* Knowledge discovery and data mining: Towards a unifying framework. In: **Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 1996. v. 96, p. 82–88. Citado na página 28.

FEKETE, J.-D.; GRINSTEIN, G.; PLAISANT, C. **IEEE InfoVis 2004 Contest – The history of InfoVis**. 2004. Citado na página 62.

FLEISHMAN, S.; COHEN-OR, D.; SILVA, C. T. Robust moving least-squares fitting with sharp features. In: ACM. **ACM Transactions on Graphics (TOG)**. [S.l.], 2005. v. 24, n. 3, p. 544–552. Citado na página 35.

FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. **Software: Practice and Experience**, v. 21, n. 11, 1991. Citado 2 vezes nas páginas 47 e 62.

GANSNER, E. R.; HU, Y.; NORTH, S. C. Visualizing Streaming Text Data with Dynamic Maps. **http://arxiv.org/abs/1206.3980**, 2012. Citado na página 47.

GÄRTNER, T. A survey of kernels for structured data. **ACM SIGKDD Explorations Newsletter**, ACM, v. 5, n. 1, p. 49–58, 2003. Citado na página 110.

GEIPEL, M. M. Self-Organization applied to dynamic network layout. **International Journal of Modern Physics C**, v. 18, n. 10, p. 1537–1549, 2007. Citado 2 vezes nas páginas 47 e 63.

GIBSON, H.; FAITH, J.; VICKERS, P. A survey of two-dimensional graph layout techniques for information visualisation. **Information visualization**, 2012. Citado 2 vezes nas páginas 47 e 63.

GOMEZ-NIETO, E.; CASACA, W.; MOTTA, D.; HARTMANN, I.; TAUBIN, G.; NONATO, L. G. Dealing with multiple requirements in geometric arrangements. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 22, n. 3, p. 1223–1235, 2016. Citado 2 vezes nas páginas 124 e 125.

GOMEZ-NIETO, E.; ROMAN, F. S.; PAGLIOSA, P.; CASACA, W.; HELOU, E. S.; OLIVEIRA, M. C. F. de; NONATO, L. G. Similarity preserving snippet-based visualization of web search results. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 20, n. 3, p. 457–470, 2014. Citado na página 27.

GRATZL, S.; LEX, A.; GEHLENBORG, N.; PFISTER, H.; STREIT, M. Lineup: Visual analysis of multi-attribute rankings. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2277–2286, 2013. Citado na página 29.

GRAUMAN, K.; DARRELL, T. The pyramid match kernel: Efficient learning with sets of features. **Journal of Machine Learning Research**, v. 8, n. Apr, p. 725–760, 2007. Citado na página 110.

GRETARSSON, B.; O'DONOVAN, J.; BOSTANDJIEV, S.; HöLLERER, T.; ASUNCION, A. U.; NEWMAN, D.; SMYTH, P. {TopicNets}: Visual Analysis of Large Text Corpora with Topic Modeling. **ACM Transactions on Intelligent Systems and Technology**, v. 3, 2012. Citado na página 46.

HAM, F. van; WATTENBERG, M.; VIEGAS, F. B. Mapping Text with Phrase Nets. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1169–1176, 2009. Citado na página 44.

HARROWER, M.; BREWER, C. a. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. **The Map Reader: Theories of Mapping Practice and Cartographic Representation**, v. 40, n. 1, p. 261–268, 2011. ISSN 0008-7041. Citado na página 83.

HAVRE, S.; SOCIETY, I. C.; HETZLER, E.; WHITNEY, P.; NOWELL, L. {ThemeRiver}: Visualizing thematic changes in large document collections. **IEEE Transaction on Visualization and Computer Graphics**, v. 8, p. 9–20, 2002. Citado 2 vezes nas páginas 43 e 44.

HINNEBURG, A. What is the nearest neighbor in high dimensional spaces? In: **Conference on Very Large Databases**. [S.l.: s.n.], 2000. Citado na página 75.

HOFFMANN, H. Kernel PCA for novelty detection. **Pattern Recognition**, v. 40, n. 3, p. 863–874, mar. 2007. ISSN 00313203. Citado 2 vezes nas páginas 107 e 116.

HOFMANN, T.; SCHÖLKOPF, B.; SMOLA, A. J. Kernel methods in machine learning. **The annals of statistics**, JSTOR, p. 1171–1220, 2008. Citado 3 vezes nas páginas 108, 109 e 110.

HONEINE, P.; RICHARD, C. Preimage problem in kernel-based machine learning. **IEEE Signal Processing Magazine**, v. 28, n. 2, p. 77–88, 2011. ISSN 10535888. Citado na página 114.

HU, Y.; WANG, Y.; WU, Y.; LI, Q.; HOU, C. Generalized Mahalanobis depth in the reproducing kernel Hilbert space. **Statistical Papers**, v. 52, n. 3, p. 511–522, 2011. ISSN 09325026. Citado 2 vezes nas páginas 115 e 116.

HU, Y. F. Efficient and high quality force-directed graph drawing. **The Mathematica Journal**, v. 10, p. 37–71, 2005. Citado 2 vezes nas páginas 47 e 62.

HUGG, J.; RAFALIN, E.; SEYBOTH, K.; SOUVAINE, D. An Experimental Study of Old and New Depth Measures. **Workshop on Algorithm Engineering and Experiments (ALENEX)**, p. 51–64, 2006. Citado 2 vezes nas páginas 123 e 125.

HYVARINEN, A. Fast and Robust Fixed-Point Algorithm for Independent Component Analysis. **IEEE Trans. Neur. Net.**, v. 10, n. 3, p. 626–634, 1999. Citado na página 86.

IZEM, R.; RAFALIN, E.; SOUVAINE, D. L. **Describing Multivariate Distributions with Nonlinear Variation Using Data Depth 1**. [S.l.], 2008. 1–20 p. Citado na página 78.

JEONG, D. H.; ZIEMKIEWICZ, C.; FISHER, B.; RIBARSKY, W.; CHANG, R. iPCA: An Interactive System for PCA based Visual Analytics. **Computer Graphics Forum**, v. 28, n. 3, p. 767–774, 2009. Citado na página 31.

JOIA, P.; PAULOVICH, F. V.; COIMBRA, D.; CUMINATO, J. A.; NONATO, L. G. Local Affine Multidimensional Projection. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2563–2571, 2011. ISSN 10772626. Citado 4 vezes nas páginas 27, 31, 86 e 98.

KASER, O.; LEMIRE, D. Tag-Cloud Drawing: Algorithms for Cloud Visualization. In: **World Wide Web Conference Commitee**. [S.l.: s.n.], 2007. Citado na página 43.

KE, Y.; SUKTHANKAR, R. Pca-sift: A more distinctive representation for local image descriptors. In: IEEE. **IEEE Computer Vision and Pattern Recognition**. [S.l.], 2004. v. 2, p. II–506. Citado na página 32.

KEIM, D. A. Visual techniques for exploring databases. In: **Knowledge Discovery in Databases**. [S.l.: s.n.], 1997. Citado 2 vezes nas páginas 29 e 30.

KEIM, D. A.; OELKE, D. Literature Fingerprinting: A New Method for Visual Literary Analysis. In: **IEEE Conference on Visual Analytics Science and Technology (VAST @ IEEE VIS)**. [S.l.: s.n.], 2007. p. 115–122. Citado na página 45.

KEIM, D. A.; WARD, M. O. Visual data mining techniques. In: BERTHOLD, M. (Ed.). **Intelligent Data Analysis: An Introduction**. Berlin: Springer, 2002. p. 2–27. Citado na página 27.

KOH, K.; LEE, B.; KIM, B.; SEO, J. ManiWordle: Providing Flexible Control over Wordle. **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 1190–1197, 2010. Citado na página 43.

KONDOR, R. I.; LAFFERTY, J. Diffusion kernels on graphs and other discrete input spaces. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2002. v. 2, p. 315–322. Citado na página 110.

KULIS, B. Metric Learning: A Survey. **Foundations and Trends in Machine Learning**, v. 5, p. 287–364, 2013. ISSN 1935-8237. Citado 4 vezes nas páginas 27, 28, 33 e 107.

KUO, B.; HENTRICH, T.; GOOD, B.; WILKINSON, M. Tag clouds for summarizing web search results. In: **World Wide Web Conference Commitee**. [S.l.: s.n.], 2007. p. 1203–1204. Citado na página 43.

LANGVILLE, A. N.; MEYER, C. D. **Google's PageRank and beyond: The science of search engine rankings**. [S.l.]: Princeton University Press, 2009. Citado na página 54.

LEE, B.; RICHE, N. H.; KARLSON, A. K.; CARPENDALE, S. SparkClouds: Visualizing Trends in Tag Clouds. **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 1182–1189, 2010. Citado na página 43.

LESPINATS, S.; AUPETIT, M. CheckViz: Sanity check and topological clues for linear and non-linear mappings. **Computer Graphics Forum**, v. 30, n. 1, p. 113–125, 2011. ISSN 01677055. Citado 3 vezes nas páginas 25, 84 e 95.

LEWIS, J.; MAATEN, L. van der; SA, V. de. A Behavorial Investigation of Dimensionality Reduction. **Proceedings of the Cognitive Science Society**, p. 671–676, 2012. Citado na página 86.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <http://archive.ics.uci.edu/ml>. Citado na página 90.

LIMPERT, E.; STAHEL, W. a.; ABBT, M. Log-normal Distributions across the Sciences: Keys and Clues. **BioScience**, v. 51, n. 5, p. 341, 2001. ISSN 0006-3568. Citado na página 89.

LIU, R. Y.; PARELIUS, J. M.; SINGH, K. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. **Annals of Statistics**, v. 27, n. 3, p. 783–858, 1999. ISSN 00905364. Citado 4 vezes nas páginas 35, 76, 83 e 86.

LIU, S.; ZHOU, M. X.; PAN, S.; SONG, Y.; QIAN, W.; CAI, W.; LIAN, X. TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis. **ACM Transactions on Intelligent Systems and Technology**, v. 3, n. 2, p. 25:1—-25:28, 2012. Citado na página 44.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of research and development**, IBM, v. 2, n. 2, p. 159–165, 1958. Citado na página 50.

LUO, D.; YANG, J.; KRSTAJIC, M.; RIBARSKY, W.; KEIM, D. A. EventRiver: Visually Exploring Text Collections with Temporal References. **IEEE Transaction on Visualization and Computer Graphics**, v. 18, n. 1, p. 93–105, 2012. Citado na página 44.

LUXBURG, U. V.; ALAMGIR, M. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2013. p. 225–233. Citado na página 75.

MAATEN, L. V. D.; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008. ISSN 02545330. Citado 2 vezes nas páginas 83 e 86.

MAIER, M.; LUXBURG, U. V.; HEIN, M. Influence of graph construction on graph-based clustering measures. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2008. p. 1025–1032. Citado na página 75.

MAO, Y.; DILLON, J. V.; LEBANON, G. Sequential Document Visualization. **IEEE Transaction on Visualization and Computer Graphics**, v. 13, n. 6, p. 1208–1215, 2007. Citado na página 46.

MARKS, J.; ANDALMAN, B.; BEARDSLEY, P. A.; FREEMAN, W.; GIBSON, S.; HODGINS, J.; KANG, T.; MIRTICH, B.; PFISTER, H.; RUML, W. *et al.* Design galleries: A general approach to setting parameters for computer graphics and animation. In: ACM PRESS/ADDISON-WESLEY PUBLISHING CO. **ACM Conference on Computer Graphics and Interactive Techniques**. [S.l.], 1997. p. 389–400. Citado na página 32.

MARTINS, R. M.; COIMBRA, D. B.; MINGHIM, R.; TELEA, A. C. Visual analysis of dimensionality reduction quality for parameterized projections. **Computers and Graphics (Pergamon)**, v. 41, n. 1, p. 26–42, 2014. ISSN 00978493. Citado na página 74.

MARTINS, R. M.; MINGHIM, R.; TELEA, A. Explaining Neighborhood Preservation for Multidimensional Projections. In: **Computer Graphics & Visual Computing (CGVC)**. [S.l.: s.n.], 2015. Citado na página 75.

MIKA, S.; SCHöLKOPF, B.; SMOLA, A.; MüLLER, K.-r.; SCHOLZ, M.; RäTSCH, G. Kernel PCA and De-Noising in Feature Spaces. **Neural Information Processing Systems (NIPS)**, v. 11, n. i, p. 536–542, 1999. ISSN 10495258. Citado na página 114.

MILLER, N. E.; Chung Wong, P.; BREWSTER, M.; FOOTE, H. Topic Islands: a wavelet-based text visualization system. In: **IEE Visualization (VIS)**. [S.l.: s.n.], 1998. p. 189–196. Citado na página 45.

MIRZARGAR, M.; WHITAKER, R. T.; KIRBY, R. M. Curve Boxplot: Generalization of Boxplot for Ensembles of Curves. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 2654–2663, dez. 2014. ISSN 1077-2626. Citado na página 83.

NEWMAN, M. E. The mathematics of networks. **The New Palgrave Encyclopedia of Economics**, Citeseer, v. 2, n. 2008, p. 1–12, 2008. Citado na página 54.

OGLIC, D.; PAURAT, D.; GäRTNER, T. Interactive Knowledge-Based Kernel PCA. In: **Machine Learning and Knowledge Discovery in Databases, ECML PKDD**. [S.l.: s.n.], 2014. p. 501–516. Citado na página 107.

OGLIC, D.; PAURAT, D.; GÄRTNER, T. Interactive knowledge-based kernel pca. In: **Machine Learning and Knowledge Discovery in Databases**. [S.l.]: Springer, 2014. p. 501–516. Citado na página 120.

OLIVEIRA, M. C. F. D.; LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 9, n. 3, p. 378–394, 2003. Citado na página 28.

PAULOVICH, F. V.; MINGHIM, R. HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1229–1236, 2008. Citado na página 46.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 2, p. 564–575, 2008. Citado 3 vezes nas páginas 31, 52 e 53.

PAULOVICH, F. V.; TOLEDO, F. M. B.; TELLES, G. P.; MINGHIM, R.; NONATO, L. G. Semantic Wordification of Document Collections. **Computer Graphics Forum**, v. 31, n. 3, p. 1145–1153, 2012. Citado 6 vezes nas páginas 14, 43, 47, 61, 65 e 66.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, 1980. Citado na página 51.

RAUBER, P. E.; FERINGA, S.; CELEBI, M. E.; TELEA, A. C.; SCIENCES, C. Interactive Image Feature Selection Aided by Dimensionality Reduction. In: **Proceedings of EuroVis Workshop on Visual Analytics**. [S.l.: s.n.], 2015. p. 2–6. Citado na página 32.

REIMANN, C.; FILZMOSER, P.; GARRETT, R.; DUTTER, R. **Statistical data analysis explained: applied environmental statistics with R**. [S.l.: s.n.], 2011. Citado na página 35.

RIECK, B.; LEITTE, H. Persistent Homology for the Evaluation of Dimensionality Reduction Schemes. **Computer Graphics Forum**, v. 34, n. 3, p. 431–440, jun. 2015. ISSN 01677055. Citado 2 vezes nas páginas 75 e 124.

SALTON, G. Developments in Automatic Text Retrieval. **Science**, v. 253, p. 974–980, 1991. Citado na página 50.

SAMMON, J. W. A Nonlinear Mapping for Data Structure Analysis. **IEEE Transactions on Computers**, C-18, n. 5, p. 401–409, 1969. ISSN 00189340. Citado na página 86.

SCHöLKOPF, B.; MIKA, S.; BURGES, C. J. C.; KNIRSCH, P.; MüLLER, K. R.; RäTSCH, G.; SMOLA, A. J. Input space versus feature space in kernel-based methods. **IEEE Transactions on Neural Networks**, v. 10, n. 5, p. 1000–1017, 1999. ISSN 10459227. Citado 4 vezes nas páginas 107, 109, 111 e 114.

SCHöLKOPF, B.; SMOLA, A.; MüLLER, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. **Neural Computation**, v. 10, n. 44, p. 1299–1319, jul. 1998. ISSN 0899-7667. Citado 6 vezes nas páginas 35, 107, 108, 109, 112 e 113.

SCHRECK, T.; LANDESBERGER, T. von; BREMM, S. Techniques for precision-based visual analysis of projected data. **Information Visualization**, Palgrave Macmillan, v. 9, n. 3, p. 181–193, 2010. ISSN 1473-8716. Citado na página 74.

SEDLMAIR, M.; MUNZNER, T.; TORY, M. Empirical guidance on scatterplot and dimension reduction technique choices. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 19, n. 12, p. 2634–2643, 2013. Citado na página 31.

SEDLMAIR, M.; TATU, A.; MUNZNER, T.; TORY, M. A Taxonomy of Visual Cluster Separation Factors. **Computer Graphics Forum**, v. 31, n. 3, p. 1335–1344, 2012. ISSN 01677055. Citado 3 vezes nas páginas 34, 35 e 125.

SEO, J.; SHNEIDERMAN, B. A rank-by-feature framework for interactive exploration of multidimensional data. **Information Visualization**, v. 4, n. 2, p. 96–113, 2005. ISSN 1473-8716. Citado na página 27.

SERFLING, R. Generalized Quantile Processes Based on Multivariate Depth Functions, with Applications in Nonparametric Multivariate Analysis. **Journal of Multivariate Analysis**, v. 83, n. 1, p. 232–247, 2002. ISSN 0047259X. Citado na página 78.

____. Depth functions in nonparametric multivariate inference. **Discrete Mathematics and Theoretical Computer Science**, American Mathematical Society, v. 72, p. 1, 2006. Citado 3 vezes nas páginas 35, 76 e 78.

SHAPIRA, L.; AVIDAN, S.; SHAMIR, A. Mode-detection via median-shift. In: **IEEE International Conference on Computer Vision**. [S.l.]: IEEE, 2009. p. 1909–1916. ISBN 978-1-4244-4420-5. Citado 2 vezes nas páginas 35 e 125.

SILVERMAN, B. W. **Density estimation for statistics and data analysis**. [S.l.]: CRC press, 1986. v. 26. Citado na página 116.

SMALL, C. G. A Survey of Multidimensional Medians. **International Statistic Review**, v. 58, n. 3, p. 263–277, 1990. Citado na página 78.

SMOLA, A. J.; KONDOR, R. Kernels and regularization on graphs. In: **Learning theory and kernel machines**. [S.l.]: Springer, 2003. p. 144–158. Citado na página 110.

SPRITZER, A. S.; Dal Sasso Freitas, C. M. Design and Evaluation of MagnetViz—A Graph Visualization Tool. **IEEE Transactions on Visualization and Computer Graphics**, v. 18, n. 5, p. 822–835, 2012. Citado na página 47.

STROBELT, H.; OELKE, D.; ROHRDANTZ, C.; STOFFEL, A.; KEIM, D. A.; DEUSSEN, O. Document Cards: A Top Trumps Visualization for Documents. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 1145–1152, 2009. Citado na página 48.

STROBELT, H.; SPICKER, M.; STOFFEL, A.; KEIM, D.; DEUSSEN, O. Rolled-out Wordles: A Heuristic Method for Overlap Removal of 2D Data Representatives. **Computer Graphics Forum**, v. 31, p. 1135–1144, 2012. Citado 4 vezes nas páginas 14, 59, 60 e 63.

SUYKENS, J.; GESTEL, T. V.; BRABANTER, J. D. **Least Squares Support Vector Machines**. World Scientific, 2002. ISBN 9789812381514. Disponível em: <http://books.google.com.br/books?id=g8wEimyEmrUC>. Citado 2 vezes nas páginas 113 e 114.

TATU, A. **Visual Analytics of Patterns in High-Dimensional Data**. Tese (Doutorado) — Universität Konstanz, 2013. Citado 4 vezes nas páginas 35, 73, 77 e 124.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On Improved Projection Techniques to Support Visual Exploration of Multidimensional Data Sets. **Information Visualization**, v. 2, n. 4, p. 218–231, 2003. Citado na página 53.

TENENBAUM, J. B. A Global Geometric Framework for Nonlinear Dimensionality Reduction. **Science**, v. 290, n. 5500, p. 2319–2323, dez. 2000. ISSN 00368075. Citado na página 90.

TURKAY, C. **Integrating Computational Tools in Interactive and Visual Methods for Enhancing High-dimensional Data and Cluster Analysis**. Tese (Doutorado) — University of Bergen, 2013. Citado na página 32.

Van Der Maaten, L.; HINTON, G. Visualizing non-metric similarities in multiple maps. **Machine Learning**, v. 87, n. 1, p. 33–55, 2012. ISSN 08856125. Citado na página 27.

VARDI, Y.; ZHANG, C.-H. The multivariate L_1 —median and associated data depth. **Proceedings of the National Academy of Sciences of the USA**, v. 97, n. 4, p. 1423–1426, 2000. ISSN 0027-8424. Citado na página 80.

VARUN, C.; ARINDAM, B.; VIPIN, K. **Anomaly detection for discrete sequences: A survey**. [S.l.], 2009. Citado na página 76.

VIéGAS, F. B.; WATTENBERG, M.; DAVE, K. Studying cooperation and conflict between authors with history flow visualizations. In: **ACM Conference on Human Factors in Computing Systems (SIGCHI)**. [S.l.: s.n.], 2004. p. 575–582. Citado na página 44.

VIéGAS, F. B.; WATTENBERG, M.; FEINBERG, J. Participatory Visualization with Wordle. **IEEE Transaction on Visualization and Computer Graphics**, v. 15, n. 6, p. 1137–1144, 2009. Citado 6 vezes nas páginas 13, 14, 40, 41, 43 e 51.

WATTENBERG, M.; VIéGAS, F. B. The Word Tree, an Interactive Visual Concordance. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1221–1228, 2008. Citado na página 44.

WILKINSON, L.; ANAND, A.; GROSSMAN, R. Graph-theoretic scagnostics. **IEEE Symposium on Information Visualization (INFOVIS @ IEEE VIS)**, p. 157–164, 2005. ISSN 1522404X. Citado 3 vezes nas páginas 14, 34 e 77.

WON, J.-H.; LIM, J.; KIM, S.-J.; RAJARATNAM, B. Condition Number Regularized Covariance Estimation. **Journal of the Royal Statistical Society. Series B, Statistical methodology**, v. 75, n. 3, p. 427–450, 2013. ISSN 1369-7412. Citado na página 81.

WU, Y.; PROVAN, T.; WEI, F.; LIU, S.; MA, K.-L. Semantic-Preserving Word Clouds by Seam Carving. **Computer Graphics Forum**, v. 30, n. 3, p. 741–750, 2011. Citado na página 43.

ZHANG, H.; Van Kaick, O.; DYER, R. Spectral mesh processing. **Computer Graphics Forum**, v. 29, n. 0, p. 1865–1894, 2010. ISSN 01677055. Citado na página 115.

ZHOU, S. K.; CHELLAPPA, R. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. **IEEE transactions on pattern analysis and machine intelligence**, IEEE INSTITUTE OF ELECTRICAL AND ELECTRONICS, v. 28, n. 6, p. 917, 2006. Citado na página 110.

ZUO, Y.; SERFLING, R. General notions of statistical depth function. **Annals of Statistics**, v. 28, n. 2, p. 461–482, 2000. ISSN 00905364. Citado na página 78.

APÊNDICE

A

PRODUÇÃO CIENTÍFICA NO PERÍODO

Lista de publicações ao longo do período do desenvolvimento desta tese:

1. CEDRIM, D., VAD, V., CASTELO, A., PAIVA, A., GRÖLLER, E., NONATO, L. G., *Depth functions as a Quality Measure and for steering Multidimensional Projections*, Computer & Graphics, 2016, DOI: 10.1016/j.cag.2016.08.008.

2. VAD, V., CEDRIM, D., BUSCH, W., FILZMOSER, P., VIOLA, I., *Generalized Box-Plot for Root Growth Ensembles*, IEEE VIS 2016 - Bio@Vis

3. SANDIM, M., CEDRIM, D., NONATO, L.G., PAGLIOSA, P., PAIVA, A., *Boundary Particle Detection Method for Particle-based Fluids*, Computer Graphics Forum, 35(2), 2016, Proceedings of Eurographics 2016;

4. VALDIVIA, P CEDRIM, D., PETRONETTO, F., PAIVA, A., NONATO, L.G., *Normal Correction., Towards Smoothing Point-based Surfaces*, XXX Sibgrapi - Conference on Graphics, Patterns and Images, Arequipa - Peru 2013;

5. PAGLIOSA, P MARTINS, R.M., CEDRIM, D., F., PAIVA, A., MINGHIM, R., NONATO, L.G., *MIST: Multiscale Information and Summaries of Texts*, XXX Sibgrapi - Conference on Graphics, Patterns and Images, Arequipa - Peru 2013;

6. CASTELO, A., CEDRIM, D., *Contagem, Enumeração e algumas Aplicações em Matemática*, 2o Colóquio de Matemática da Região Sudeste, 2013