# Atelier Visualisation d'informations, Interaction et Fouille de données

Organisateurs : : Pierrick Bruneau (LIST), Mohammad Ghoniem (LIST), Fabien Picarougne (LS2N)

# PRÉFACE

Désormais bien établi à EGC, l'atelier VIF émane du groupe de travail *Visualisation d'Information, Interaction et Fouille de données*, fruit de la collaboration entre les associations EGC et AFIHM. Celui-ci se propose de faire le point sur l'actualité en visualisation interactive d'informations, tant du point de vue fondamental qu'applicatif. Á la confluence des communautés EGC et VIS et à la croisée des disciplines (Informatique, Géographie, Ergonomie, Design, etc.), les méthodes de visualisation interactive et de fouille visuelle des données sont au cœur des préoccupations de cet atelier. Aussi, il aura pour vocation de favoriser l'échange sur l'évolution récente des axes de recherche dans ces thématiques, et sur l'application des méthodes de visualisation à des problématiques industrielles. Le traitement de données massives (*Big Data*) et des flux de données fera l'objet d'une attention particulière.

Pierrick Bruneau
LIST

Mohammad Ghoniem
LIST

Fabien Picarougne
LS2N

**Membres du comité de lecture**

Le Comité de Lecture est constitué de:

Michaël Aupetit (Qatar Computing Research Institute)
Hanene Azzag (LIPN, Université de Paris 13 Sorbonne)
David Bihanic (Université Paris 1 Panthéon-Sorbonne)
Fatma Bouali (LI Tours et Université de Lille 2)
Pierrick Bruneau (Luxembourg Institute of Science and Technology)
Mohammad Ghoniem (Luxembourg Institute of Science and Technology)
Fabrice Guillet (LINA, Université de Nantes)
Patrik Hitzelberger (Luxembourg Institute of Science and Technology)
Pascale Kuntz (LINA, Université de Nantes)
Nicolas Labroche (LI, Université François Rabelais de Tours)
Guy Mélançon (LABRI, Université de Bordeaux)
Monique Noirhomme (Institut d'Informatique, FUNDP, Namur, Belgique)
Benoit Otjacques (Luxembourg Institute of Science and Technology)
Fabien Picarougne (LINA, Université de Nantes)
Bruno Pinaud (LABRI, Université de Bordeaux)
Julien Velcin (Université de Lyon 2)
Gilles Venturini (LI, Université François Rabelais de Tours)

## TABLE DES MATIÈRES

# Layout radial 3D pour la visualisation de la centralité dans les graphes

Piriziwè Kobina*, Thierry Duval* Laurent Brisson*

*IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
{piriziwe.kobina, thierry.duval, laurent.brisson}@imt-atlantique.fr
http://www.imt-atlantique.fr

## 1 Introduction

Freeman (2004) définit l'analyse des réseaux sociaux comme une méthodologie d'étude des relations entre acteurs sociaux. Elle permet de modéliser et de visualiser ces réseaux d'acteurs à l'aide de graphes où on peut souvent trouver des regroupements d'acteurs suivant une certaine affinité (Moreno, 1953) et des acteurs isolés. Par contre, l'interprétation de ce sociogramme d'affinités et/ou de rejets est parfois difficile (Lancichinetti et al., 2010; Dao et al., 2017). Il est donc courant de déterminer des métriques qui permettent de caractériser les affinités ou les rejets pouvant être observés ou encore l'importance de chaque acteur dans le réseau, et trouver une technique de visualisation qui permet de mettre en exergue ces métriques.

Pour cela, de nombreuses techniques de visualisation de graphes 2D sont utiles pour visualiser les affinités dans les réseaux ou pour visualiser l'importance ou le rôle que joue chaque acteur dans le réseau. Par contre, face à des données volumineuses et complexes, ces techniques sont généralement dans l'incapacité de fournir une visualisation appropriée, par manque d'espace d'affichage par exemple. Par conséquent, l'analyse devient compliquée. Il est donc nécessaire d'augmenter l'espace d'affichage de données, et pour cela il est possible d'adapter certaines techniques 2D à la 3D (Brisson et al., 2018; Cliquet et al., 2017), ce qui reste encore un vaste domaine à explorer (Spritzer et Freitas, 2008).

Dans cet article, nous illustrons nos contributions du passage à la 3D d'une technique 2D qui permet de mettre en évidence l'importance des nœuds dans le graphe par leur centralité.

## 2 Visualisation 2D de la centralité

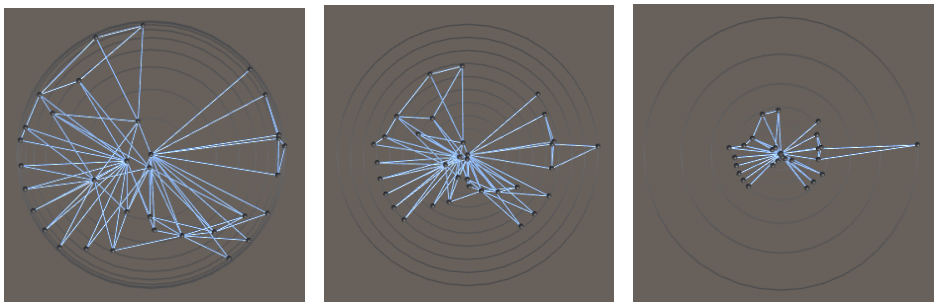
Dans un graphe, on aurait besoin de trouver les nœuds les plus importants c'est-à-dire les nœuds qui font le lien entre les autres ou les nœuds les plus centraux, etc.

Ainsi, l'importance d'un nœud dans un graphe dépend de l'intérêt métier et peut être caractérisée par un certain nombre de métriques telles que les mesures de centralité. Dans nos travaux, nous nous intéressons à la centralité de proximité qui montre à quel point un nœud est proche de tous les autres dans le graphe (Freeman, 1978).

Les premiers travaux de Brandes et al. (2003), sur la visualisation de graphes, permettent de montrer la mise en évidence de la centralité de proximité. Brandes et Pich (2011) proposent

ensuite une approche radiale 2D en matérialisant la notion de centralité par des cercles concentriques afin d'illustrer l'importance des nœuds dans le graphe. Cette approche est basée sur une extension de l'algorithme de minimisation des contraintes (MDS) en incluant des rayons de cercles déterminés à partir des valeurs de centralité des nœuds. Pour cela, les nœuds ayant une centralité forte sont au centre et ceux de faible centralité à la périphérie. Ils proposent également de mettre l'emphase sur le centre du réseau (Figure 1a) ou bien sur sa périphérie (Figure 1c). L'emphase sur le centre consiste à rendre plus distinctifs les nœuds qui sont au centre des cercles, ce qui concentre tous les autres nœuds à la périphérie. Par contre, l'emphase sur la périphérie étale les nœuds de la périphérie.

Ainsi, la figure 1 montre la mise en évidence de la centralité de proximité avec le fameux réseau du club de karaté étudié par Zachary (1977). Ce réseau décrit 78 relations d'amitié entre 34 membres du club de karaté d'une université américaine dans les années 70.



(a) emphase sur le centre (b) vue uniforme (c) emphase sur la périphérie

FIG. 1: Visualisation radiale du club de karaté de Zakary

# 3 Extension 3D de la visualisation radiale de graphes

Pour étendre la représentation radiale uniforme au domaine 3D, nous proposons de la projeter, selon l'axe vertical, sur trois surfaces 3D différentes de façon à toujours garder la vue radiale 2D en vue de dessus. Les trois surfaces de projection sont : une demi-sphère, un cône et une portion de tore. Ainsi, l'ajout d'une troisième dimension à la représentation radiale 2D permet de mieux distinguer la connectivité des nœuds, car on obtient une répartition verticale.

## 3.1 Extension 3D de la vue uniforme

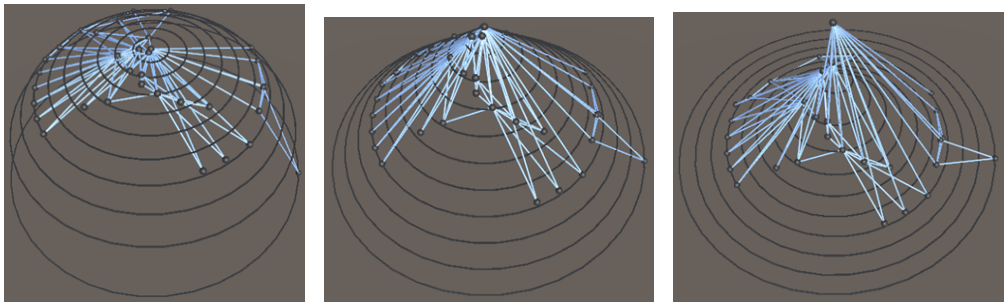
Du point de vue de la répartition des nœuds selon l'axe vertical, la projection sphérique (Figure 2a) concentre les nœuds centraux et étale les nœuds en périphérie. La projection conique (Figure 2b) répartit, quant à elle, les nœuds uniformément alors que celle sur la portion de tore (Figure 2c) étale les nœuds centraux et écrase ceux en périphérie.

Lorsqu'on change l'angle de vue de la projection sphérique, on a la possibilité de voir les nœuds en périphérie. Ainsi, une élévation uniforme sur la demi-sphère (2a) peut fournir à la fois les avantages de deux représentations 2D : la vue uniforme (Figure 1b) et l'emphase sur la

périphérie (Figure 1c). Par contre, on ne peut pas voir les nœuds centraux, du fait que ceux-ci soient sur la partie supérieure de la demi-sphère. De plus, avec la projection sphérique, les liens entre les nœuds centraux et ceux en périphérie sont à l'intérieur de la surface de projection.

La projection conique regroupe à la fois les avantages des trois représentations 2D : vue uniforme, emphase sur le centre et emphase sur la périphérie. En effet, elle fournit modérément une visualisation des nœuds centraux et des nœuds en périphérie. Avec cette projection, les liens sont majoritairement sur la surface de projection. Aussi, les liens sont beaucoup lisibles entre les nœuds centraux et les nœuds en périphérie, comparativement à l'approche sphérique.

En ce qui concerne la projection sur la portion de tore, on a une excellente vue sur les liens entre le centre et la périphérie, comparativement à la projection conique, car avec cette projection, les liens sont à l'extérieur de la surface de projection. Aussi, elle permet distinguer les nœuds centraux sur le dessus de la surface et les nœuds en périphérie.



(a) Projection sphérique (b) Projection conique (c) Projection sur la portion de tore

FIG. 2: Visualisation radiale uniforme

### 3.2 Discussion

Contrairement à la représentation 2D, une élévation uniforme permet d'améliorer nettement la perception de la connectivité des nœuds. En effet, elle fournit une excellente visualisation des nœuds centraux et des nœuds en périphérie. Aussi, on a une meilleure vue des liens entre le centre et la périphérie en fonction de la surface.

Néanmoins, il existe des recouvrements de liens dans les zones denses suivant la surface de projection. En effet, avec les projections sphérique et conique, les liens entre les nœuds sont respectivement à l'intérieur de la surface et sur la surface et certains liens sont masqués par d'autres. Par contre, sur la portion de tore, les liens sont à l'extérieur de la surface et il y a moins de recouvrement comparativement aux approches sphérique et conique.

## 4 Conclusion et perspectives

Les résultats de nos travaux sont une première proposition du passage à la 3D du concept de "layout radial". Ils nous ont permis de montrer qu'une élévation sphérique donne les mêmes avantages que l'emphase périphérique ; une élévation sur une portion de tore fournit les avantages des emphases centrale et périphérique et une élévation sur un cône les avantages des

emphases centrale et périphérique, mais moins bien que la démi-sphère et la portion de tore. Aussi, cette élévation améliore la perception de liens entre les nœuds.

Comme perspectives, nous étudierons les résultats du passage en 3D des représentations mettant déjà l'emphase sur le centre et sur la périphérie. Aussi, par manquent d'espace d'affichage, certains nœuds et liens seraient couverts par d'autres avec la représentation 2D face à une taille importante de nœuds et liens. Ainsi, nous traiterons des données plus volumineuses en terme de nombre de nœuds et de liens.

# Références

Brandes, U., P. Kenis, et D. Wagner (2003). Communicating centrality in policy network drawings. *IEEE Transactions on Visualization and Computer Graphics 9*, 241–253.

Brandes, U. et C. Pich (2011). More flexible radial layout. *Journal of Graph Algorithms and Applications 15*, 151–173.

Brisson, L., T. Duval, et R. Sahl (2018). Visualisation immersive de graphes en 3D pour explorer des graphes de communautés. *EGC-VIF*.

Cliquet, G., M. Perreira, F. Picarougne, Y. Prié, et T. Vigier (2017). Towards HMD-based Immersive Analytics. In *Immersive Analytics workshop of IEEE VIS 2017*.

Dao, V.-L., C. Bothorel, et P. Lenca (2017). Community structures evaluation in complex networks: A descriptive approach. In E. Shmueli, B. Barzel, et R. Puzis (Eds.), *3rd International Winter School and Conference on Network Science*, pp. 11–19. Springer International Publishing.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 215–239.

Freeman, L. C. (2004). The development of social network analysis.

Lancichinetti, A., M. Kivelä, J. Saramäki, et S. Fortunato (2010). Characterizing the community structure of complex networks. *PLoS ONE 5*.

Moreno, J. L. (1953). *Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama, 2nd ed.* Oxford, England: Beacon House.

Spritzer, A. et C. Freitas (2008). Navigation and interaction in graph visualizations. *Revista de Informática Teórica e Aplicada 15*, 111–136.

Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research 33*, 452–473.

# Summary

This paper presents new methods of 3D visualization of graphs: projections on a half-sphere, a cone and a portion of torus. These methods allow to illustrate the importance of nodes in the graph in terms of centrality.

# Visualisation interactive des graphes d'une règle d'association informative

## 1 Introduction et Motivations

Ce papier aborde la question de visualisation interactive des graphes d'une règle d'association $X \rightarrow Y$. De telle règle, $X$ est la prémisse et $Y$ le conséquent, tels que $X \cap Y = \emptyset$. Une règle d'association est apprise à partir d'une base de données $\mathcal{B} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$, où $\mathcal{I}$ est l'ensemble de motifs, $\mathcal{T}$ l'ensemble de transactions, et $\mathcal{R}$ une relation binaire entre $\mathcal{T}$ et $\mathcal{I}$. Le calcul de telle règle à partir du couple support-confiance (Agrawal et Srikant, 1994) est un exemple classique de la littérature. Toutefois, ce couple support-confiance ne porte pas sur les règles négatives [1] pouvant être une source d'informations pertinentes, mais plutôt sur les règles positives dont plusieurs sont inintéressantes et redondantes qui viennent bruiter les graphes.

## 2 Contributions

Afin de pallier les limites signalées, nous proposons une méthodologie de visualisation très lisible, interactive et interprétable des graphes des règles positives et négatives les plus informatives à l'aide de la mesure la plus sélective, $M_{GK}$ (Feno, 2007). Une règle *informative* est celle qui, à partir de la plus petite prémisse, fournit le plus grand conséquent. Pour ce faire, nous avons exploité les concepts d'un générateur et d'un fermé en utilisant le couple support-$M_{GK}$. Les applications $t : \mathcal{I} \mapsto \mathcal{T}, t(X) = \{y \in \mathcal{T} | \forall x \in \mathcal{I}, x\mathcal{R}y\}$ et $i : \mathcal{T} \mapsto \mathcal{I}, i(Y) = \{x \in \mathcal{I} | \forall y \in \mathcal{T}, x\mathcal{R}y\}$ définissent la connexion de Galois, telles que $\gamma(X) = iot(X) = i(t(X))$ s'appelle l'opérateur de fermeture. $X$ est dit fermé si $X = \gamma(X)$, et est générateur s'il est minimal dans sa classe d'équivalence : $[X] = \{Y \subseteq \mathcal{I} | \gamma(Y) = \gamma(X)\}$. Le support, la confiance et la $M_{GK}$ de la règle $X \rightarrow Y$ s'écrivent respectivement $supp(X \cup Y) = \frac{|t(X \cup Y)|}{|\mathcal{T}|}$, $P(Y|X) = \frac{supp(X \cup Y)}{supp(X)}$ et $M_{GK}(X \rightarrow Y) = \frac{P(Y|X) - P(Y)}{1 - P(Y)}$ si $P(Y|X) > P(Y)$ et $M_{GK}(X \rightarrow Y) = \frac{P(Y|X) - P(Y)}{P(Y)}$ sinon, où $P(A) = supp(A)$ est la probabilité d'apparition de $A$. On entend d'une règle *inintéressante* que son intensité $M_{GK}$ est négative ou nulle. Autrement dit, $X \rightarrow Y$ est inintéressante si l'apparition de $X$ diminue les chances d'apparition de $Y$. Afin d'élaguer ce type inintéressant, nous avons comparé

1. C'est une implication logique de la forme $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$ et $\overline{X} \rightarrow \overline{Y}$, $\forall X, Y \subseteq \mathcal{I}$, où $\overline{I} = \neg I = \mathcal{I} \backslash I$.

la confiance $P(Y|X)$ et la probabilité du conséquent $P(Y)$. En effet, si $P(Y|X) > P(Y)$, alors $X \rightarrow Y$ est potentiellement intéressante. Si $P(Y|X) \leq P(Y)$, alors $X \rightarrow Y$ est inintéressante. Grâce à ces propriétés sémantiques, les règles d'association inintéressantes étant systématiquenent élaguées. Il reste à étudier dans ce qui suit les redondances. Rappelons que la règle $r_1 : X_1 \rightarrow Y_1$ est redondante moins informative par rapport à la règle $r_2 : X_2 \rightarrow Y_2$ ssi : $(i)$ $(supp(r_1) = supp(r_2)$ et $M_{GK}(r_1) = M_{GK}(r_2))$, $(ii)$ $(X_2 \subseteq X_1$ et $Y_1 \subset Y_2)$. De telle formalisation, extraire simultanément les règles d'association positives et négatives augmente considérablement les bruits et la complexité de calcul, ce qui va restaurer 8 règles ($X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$, $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ et $Y \rightarrow \overline{X}$). A cela, différentes propriétés d'élagage de l'espace de recherche ont été définies. A cet effet, nous n'avons retenu que la moitié de l'ensemble ($X \rightarrow Y$, $\overline{X} \rightarrow \overline{Y}$, $X \rightarrow \overline{Y}$ et $\overline{X} \rightarrow Y$) et n'avons étudié que 2 règles ($X \rightarrow Y$ et $X \rightarrow \overline{Y}$), soit un taux de réduction 75% de l'espace de recherche.

Relativement à ces deux types, nous avons développé 4 nouvelles bases. La 1re concerne la base de règles positives exactes (i.e. $M_{GK}(X \rightarrow Y) = 1$), qui sélectionne les prémisses (resp. conséquents) dans des générateurs (resp. fermés). La 2e est la base de règles positives approximatives (i.e. $M_{GK}(X \rightarrow Y) < 1$), où les premisses sont sélectionnées dans des générateurs, et les conséquents dans d'autres fermés contenant le fermé courant d'un générateur. La 3e base est la base de règles négatives exactes (i.e. $M_{GK}(X \rightarrow \overline{Y}) = 1$), qui sélectionne les prémisses dans des générateurs d'un itemset fréquent maximal de la *bordure positive* $Bd^+(\mathcal{F})$, et les conséquents dans des traverses minimales, notées $\widetilde{Bd^+(\mathcal{F})}$, c'est-à-dire dualisation de $Bd^+(\mathcal{F})$, où $\mathcal{F}$ est l'ensemble des motifs fréquents de la base de données $\mathcal{B}$. On entend d'un motif fréquent que son support dépasse un seuil minimum de support, $minsup \in [0, 1]$. La 4e base est la base de règles négatives approximatives (i.e. $M_{GK}(X \rightarrow \overline{Y}) < 1$), qui sélectionne les prémisses et conséquents dans des générateurs des fermés respectivement incomparables.

Nous avons implémenté notre approche avec l'outil `rchicmgk`, sous forme d'un package interfaçage R. L'idée est d'offrir un moyen visuel aux experts pour interagir et comprendre les phénomènes sous-jacents. À travers ces visualisations opérationnelles, on peut supprimer (resp. changer) temporairement les motifs (resp. seuils de test), puis l'outil met à jour à nouveau les graphes sans re-importer les données. En dépit de sa simplicité, ce travail rend aux experts un outil puissant pour la fouille de données massives, et permet d'aider ces experts dans leur prise de décision. Les perspectives sont nombreuses. Actuellement, l'étiquetage typologique des règles retenues peut se faire à la main mais pourrait très bien s'automatiser. L'extension de ce travail aux paradigmes des règles multidimensionnelles pourrait être aussi envisageable.

# Références

Agrawal, R. et R. Srikant (1994). Fast Algorithms for Mining Association Rules. In *Proceedings of 20th VLDB Conference*, Santiago, Chile, pp. 487–499.

Feno, D. (2007). *Mesures de qualité des règles d'association : normalisation et caractérisation de bases*. Ph. D. thesis, Université de La Réunion, France.

# Summary

We proposed a methodology for interactive visualization of the graphs of an association rule. We have implemented this approach with package `rchicmgk`, interface R.

# Aider l'utilisateur à mieux visualiser

Barthélémy Serres[*,**], Fatma Bouali[**,***]
Gilles Venturini[**,*]

[*]ILIAD3, Université de Tours, 64 avenue Jean Portalis, 37200 Tours, France
barthelemy.serres@univ-tours.fr,
http://iliad3.univ-tours.fr
[**]LIFAT, EA 6300, Université de Tours, 64 avenue Jean Portalis, 37200 Tours, France
venturini@univ-tours.fr
http://lifat.univ-tours.fr
[***]Université de Lille, IUT C, 53 Rue de l'Alma, BP 557, 59100 - Roubaix Cedex, France
fatma.bouali@univ-lille.fr

## 1 Introduction

Dans les approches centrées utilisateurs en fouille de données, on considère que les moyens visuels et interactifs mis en oeuvre vont faciliter le processus d'exploration des données et des connaissances, et vont , du fait de leur nature centrée utilisateur, résoudre certains des problèmes inhérents aux outils classiques ne faisant pas intervenir l'utilisateur. Ainsi les visualisations sont en principe plus faciles à interpréter que des sorties d'algorithmes non interactifs, et doivent permettre par exemple une meilleure validation des résultats. Les interactions et les requêtes graphiques présentes dans les visualisations sont supposées être, par définition, plus simples à utiliser et à maîtriser que des requêtes standards (dans un langage d'interrogation, sous forme de scripts etc). Cependant, dire que les outils visuels et interactifs sont faciles à utiliser par les utilisateurs est un raccourci qui omet la complexité inhérente des visualisations et la difficulté qu'il y a à les choisir et à la configurer. Pour des utilisateurs novices, trouver la bonne visualisation (et ses interactions), avec le bon choix des attributs de données et de leur affectation à des attributs visuels, peut poser problème. Dans cette présentation, nous allons observer trois catégories de travaux qui tentent de résoudre certains des problèmes qui se posent dans la thématique globale de l'aide que l'on peut proposer aux utilisateurs de visualisations. Ces trois catégories sont 1) les assistants utilisateurs, 2) les outils de gestion d'historique de visualisation, 3) les approches posant directement ce problème sous la forme d'un problème d'optimisation d'une fonction.

Pour commencer nous allons souligner tout au long de la présentation des dénominateurs communs à ces trois approches. Trouver la bonne visualisation (les bons attributs visuels, leur configuration) consiste à trouver parmi un ensemble possible de configurations celle qui pourra résoudre au mieux le problème posé par l'utilisateur. Cette notion d'ensemble des configurations peut être formalisée comme un espace de recherche, dans lequel il faut trouver les points intéressants. Ensuite, les trois approches étudiées peuvent être vue comme des mécanismes de

recherche. Les assistants sont des heuristiques qui vont sélectionner certains points de l'espace de recherche. Les approches à base d'historique permettent de représenter le cheminement de l'utilisateur dans l'espace de recherche, et laisse l'utilisateur revenir en tout point de cette arborescence. Les systèmes réalisant une optimisation des visualisations utilisent des opérateurs de recherche soit exhaustifs (toutes les possibilités sont explorées), soit des méta-heuristiques ou d'autre approches. Enfin, même si cela peut dans un premier temps paraître paradoxal, certains des problèmes de visualisation peuvent accepter la définition d'une fonction à optimiser. Le paradoxe vient du fait que précisément, un des reproches que l'on fait aux méthodes classiques est que les critères mathématiques ne correspondent pas à ce que l'expert a en tête. Dans le cas de l'optimisation des visualisations, le problème est quelque peu transposé : on cherche une visualisation qui va maximiser une forme d'efficacité visuelle.

Parmi les assistants utilisateurs, nous distinguons trois approches : à base de connaissances (les assistants vont utiliser des régles et "lois" connues en visualisation pour proposer la meilleure visualisation possible et son paramétrage) (Mackinlay, 1986), à base de comportement utilisateur (à partir des choix effectués par les utilisateurs précédents, des recommandations sont effectuées) (Key et al., 2012).

En ce qui concerne les systèmes de gestion d'historique (Bavoil et al., 2005), nous citerons quelques exemples de systèmes et comment ils visualisent les évolutions que l'utilisateur peut apporter aux visualisations.

Enfin, nous citerons plusieurs exemples d'optimisation en visualisation (Bertini et al., 2011) (Liiv, 2010) (Fruchterman et Reingold, 1991) : la réorganisation de lignes/colonnes (exemple classique et ancien avec en notamment les travaux de Bertin), le dessin de graphes, etc. Nous ferons un zoom sur ScagExplorer (Dang et Wilkinson, 2014) et la manière dont il explore et évalue tous les couples d'axes possibles, avec au final une représentation de clusters de graphiques. Nous aborderons par ce biais la question de la définition de fonctions permettant de quantifier la qualité d'une visualisation (Albuquerque et al., 2011) (Pineo et Ware, 2011).

En conclusion, nous soulignons que les approches visuelles et interactives, centrées utilisateur et par principe proches de lui, doivent le rester et ne pas négliger l'aide qu'il faut apporter à l'utilisateur pour trouver la bonne visualisation et le paramétrage qui va avec.

# Références

Albuquerque, G., M. Eisemann, et M. Magnor (2011). Perception-based visual quality measures. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pp. 13–20. IEEE.

Bavoil, L., S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, et H. T. Vo (2005). Vistrails : Enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.*, pp. 135–142. IEEE.

Bertini, E., A. Tatu, et D. Keim (2011). Quality metrics in high-dimensional data visualization : An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics 17*(12), 2203–2212.

Dang, T. N. et L. Wilkinson (2014). Scagexplorer : Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific visualization symposium*, pp. 73–80. IEEE.

Fruchterman, T. M. et E. M. Reingold (1991). Graph drawing by force-directed placement. *Software : Practice and experience 21*(11), 1129–1164.

Key, A., B. Howe, D. Perry, et C. Aragon (2012). Vizdeck : self-organizing dashboards for visual analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 681–684. ACM.

Liiv, I. (2010). Seriation and matrix reordering methods : An historical overview. *Statistical Analysis and Data Mining : The ASA Data Science Journal 3*(2), 70–91.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Trans. Graph. 5*(2), 110–141.

Pineo, D. et C. Ware (2011). Data visualization optimization via computational modeling of perception. *IEEE transactions on visualization and computer graphics 18*(2), 309–320.

# Summary

This paper proposes a discussion about the help that can be provided to users who have visualization needs. We present three possible axes for helping users: 1) user assistants that try to automatically provide and configure visualizations that match the user data and objectives, 2) history mechanism, as a way to help users find a relevant visual representation of the data, 3) the optimization of visualizations, in which an objective function is defined as well as a search space of visualizations. We argue that these approaches can be modeled as search procedures that look for efficient solutions in a search space of visualizations.

# Combiner arbres phylogénétiques et visualisation d'ensembles

Jean-Baptiste Lamy[*,**], Flora Jay[*,***]

[*]Laboratoire de Recherche en Informatique,
CNRS/Université Paris-Sud/Université Paris-Saclay, Orsay, France
flora.jay@lri.fr
[**]LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny,
INSERM UMRS 1142, Sorbonne Universités
jean-baptiste.lamy@univ-paris13.fr
[***]Laboratoire EcoAnthropologie et Ethnobiologie,
CNRS/MNHN/Université Paris Diderot, Paris, France

Les arbres sont très largement utilisés en phylogénie. Cependant, un arbre avec $n$ feuilles présente les similarités entre $n-1$ sous-ensembles de feuilles, alors qu'il existe $2^n - n - 1$ sous-ensembles possibles d'au moins deux feuilles. Par exemple, pour 3 feuilles $A$, $B$ et $C$, 4 similarités peuvent être mesurées, entre les sous-ensembles $\{A, B\}$, $\{B, C\}$, $\{A, C\}$ et $\{A, B, C\}$, mais un arbre n'en montrera que 2, *e.g.* $\{A, B\}$ et $\{A, B, C\}$ si une branche rassemble $A$ et $B$. Plus $n$ augmente, plus la vision donnée par l'arbre deviendra réductrice.

Ce problème est particulièrement important en génétique des populations (Pickrell et Pritchard, 2012), lorsque l'on étudie la diversité génétique des populations d'êtres vivants et leurs relations. Dans ce contexte, il est nécessaire de tenir compte des processus biologiques telle l'apparition de mutations dans le génome mais aussi des processus démographiques, tels que la séparation des populations (aussi appelée divergence), la variation de leur taille effective, et les migrations (mouvement d'un groupe d'individus qui quittent une population pour en rejoindre ou en créer une autre, apportant au passage leur matériel génétique). Un arbre généalogique peut représenter la composante évolutive et une partie de la composante démographique (divergence simple des populations, dérive génétique plus forte dans les populations de petite taille, etc) mais pas la composante migratoire post divergence qui induit des "flux de gènes" entre branches de l'arbre.

Afin de résoudre ce problème, nous proposons l'utilisation de visualisation d'ensembles, et notamment des boîtes arc-en-ciel (Lamy et al., 2017) et de leur variante proportionnelle (Lamy et Tsopra, 2019), et l'illustrons par une application à un sous-ensemble de données extraites du *1000 Genomes Project* (Auton et al., 2015). Une première approche consiste à visualiser les similarités comme des ensembles (sous-ensemble des populations avec les mêmes mutations).

Une seconde approche consiste à superposer arbre phylogénétique et boîtes arc-en-ciel (voir Figure 1). L'arbre (en noir) représente l'histoire démographique "prépondérante" des populations, lesquelles sont présentées en colonne (les couleurs dans les en-têtes de colonne identifient les continents). La longueur des branches de l'arbre indique le nombre de mutations depuis l'ancêtre commun. Les boîtes rectangulaires permettent de visualiser les similarités entre branches éloignées de l'arbre : par exemple la boîte bleu ciel à gauche montre une similarité

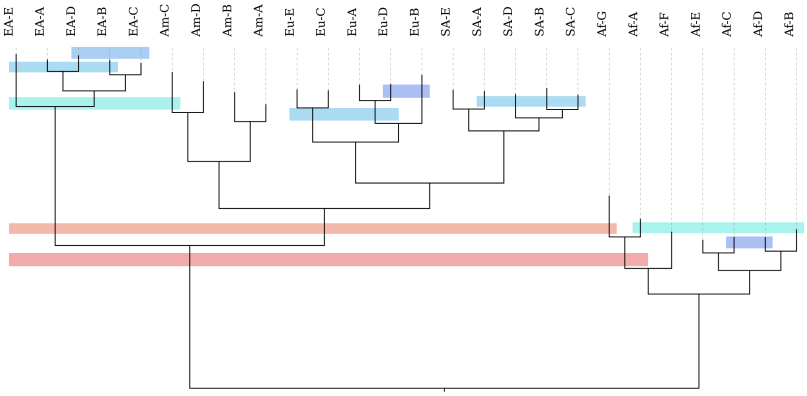



FIG. 1 – *Exemple préliminaire de combinaison arbre phylogénétique - boîtes arc-en-ciel.*

entre une population amérindienne (Am-C) et l'ancêtre des populations est-asiatique (EA), bien qu'il n'y ait pas d'ancêtre commun spécifique à l'ensemble de ces populations. Chaque boîte recouvre les branches (et sous-branches) de l'arbre correspondant aux populations qui partagent les similarités. La hauteur des boîtes est proportionnelle au nombre de mutations similaires, et la couleur indique le nombre de populations impliquées (plus la couleur chaude, plus le nombre de populations est élevé). Les boîtes correspondant à des similarités déjà visualisées par l'arbre ne sont pas affichées. Ces analyses se basent sur un sous-ensemble de marqueurs génétiques et nous demeurons prudents quant à leur interprétation.

Les difficultés rencontrées sont (1) la génération d'arbres et de boîtes arc-en-ciel ayant la même unité afin de les rendre comparables, (2) la génération de boîtes à partir de données numériques (une mutation donnée pouvant être présente chez 60% des individus d'une population, et non seulement 0 ou 100%), et (3) la complexité visuelle des représentations obtenues.

# Références

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, et et al. (2015). 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature 526*(7571), 68–74.

Lamy, J. B., H. Berthelot, C. Capron, et M. Favre (2017). Rainbow boxes : a new technique for overlapping set visualization and two applications in the biomedical domain. *Journal of Visual Language and Computing 43*, 71–82.

Lamy, J. B. et R. Tsopra (2019). RainBio : Proportional visualization of large sets in biology. *IEEE Transactions on Visualisation and Computer Graphics accepted*.

Pickrell, J. K. et J. K. Pritchard (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics 8*(11), e1002967.

# Summary

In population genetics, it is important to visualize both the genetic evolution of populations and the contribution of migrations to the spreading of genetic mutations. Here, we propose an original approach combining phylogenetic tree and set visualization with rainbow boxes.

# Index