## **Visualizing symmetric square matrices with rainbow boxes:**

## Methods and application to character co-occurrence matrices in literary texts

Jean-Baptiste Lamy

jean-baptiste.lamy @ univ-paris13.fr



LIMICS Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny Sorbonne Universités, Paris INSERM UMRS 1142

# Introduction

#### Symmetric square matrices / co-occurrence matrices [Leydesdorff]

- Frequent type of dataset
  - Example: distant reading in digital humanities [Jänicke]
  - Character matrix in a novel

				С					SS	0			ur		ne		_				
		uu		apitan	e		kang	am	antelle	asinne	am	SS	mine		usculi	aline	Saphi	ald	ed	Ъ	
	Tienn	Rake	Alyse	Le C	Gauv	Méric	Rash	Chéra	Lesc	La Ri	Werfa	Augu	Le La	Nicol	Crépi	Auroi	King	Théra	Malfr	Mabii	Éraro
Tienn		1	1	1	1	0	0	1	0	2	2	2	2	5	3	0	0	0	4	4	4
Rakenn			1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alyse				1	1	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Le Capitann					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gauve						2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Méric							2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Rashkang								2	2	0	0	0	0	0	0	0	0	0	0	0	0
Chéram									1	0	0	0	0	0	0	0	0	0	0	0	0
Lescantelles										1	1	0	0	0	0	0	0	0	0	0	0
La Rasinne											1	1	1	5	0	0	0	0	0	0	0
Werfam												1	1	5	0	0	1	0	0	0	0
Auguss													1	5	0	0	0	0	0	0	0
Le Lamineur														5	0	0	0	0	0	0	0
Nicol															0	0	0	0	0	0	0
Crépusculine																1	1	1	0	0	0
Auroraline																	1	1	0	0	0
King Saphir																		1	0	0	0
Thérald																			0	0	0
Malfred																				4	4
Mabine																					4
Érard																					

# **Matrix-based visualization**

Example on
 *'Les Misérables'* (Victor Hugo)

### Problem

 A character can belong to at most one group / cluster





#### Propose a visualization techniques for symmetric square matrices

- By transforming the matrix into overlapping sets
- And visualizing these overlapping sets with rainbow boxes
- Focus on the representation of subsets of interrelated elements
  - For example in a novel: groups of characters that know each other

#### Rainbow boxes have been used only in the biomedical domain

=> a new application domain



# **Rainbow boxes**

#### **Example on amino acids**

Е	D	Р	Ν	S	Α	G	С	Т	V	I	L	М	F	W	Y	Н	К	R	Q
Nega	ative			Tiny					Alipl	natic						Posi	tive		
	Sma	.11									1 1 1 1 1	1 1 1 1 1	Aron	natic					
					Hydr	ropho	bic												
											1 1 1 1	1 1 1 1		Pola	r				

# **General methods**

**A symmetric square matrix**  $M = (M_{i,j})_{1 \le i \le n, 1 \le j \le n} \in \mathbb{R}$ , with  $M_{i,j} = M_{j,i}$ 

#### Extracting overlapping sets

- One element for each row / column
- One set for each group of interrelated elements

#### The selection function

 Returns True if two elements are related, depending on the value in the matrix

 $select: \mathbb{R} \to \{True, False\}$ 

Compute S, the set of sets:

$$S_0 = \left\{ s \subseteq E \mid |s| > 1 \land \forall (i,j) \in (s,s) \text{ with } i \neq j, select(M_{i,j}) = True \right\}$$
$$S = \left\{ s \in S_0 \mid \not\exists s' \in S_0, s \subset s' \right\}$$

#### One group



# **General methods**

Each set corresponds to several values in the matrix

The aggregation function returns a single value from these values:

 $aggregate: \mathbb{R}^p \to \mathbb{R}, \text{ with } p \geq 1$ 



- **The colorization function converts the resulting value to a color:**  $colorize : \mathbb{R} \rightarrow color$
- Each set => One rectangular box in rainbow boxes

Visualization parameters: the 3 functions

select(), aggregate(), colorize()

Matrix produced manually by the authors

### Number in the matrix: part of the novel in which the two characters begin their relation

♦ 0: never

- ♦ 1: before the novel begins
- ♦ 2-5: parts 1-4

#### **Definitions of the 3 functions:**

 $select: p \mapsto True \ if \ p \neq 0, False \ otherwise$  $aggregate: p_1, p_2, \dots \mapsto \min(p_1, p_2, \dots)$  $\begin{cases} violet \quad if \ a = 1\\ cyan \quad if \ a = 2\\ green \quad if \ a = 3\\ yellow \quad if \ a = 4\\ red \quad if \ a = 5 \end{cases}$ 

				nn			0		lles	ре			eur		lline	e	hir					
	Tienn	Rakenn	Alyse	Le Capita	Gauve	Méric	Rashkan	Chéram	Lescante	La Rasin	Werfam	Auguss	Le Lamin	Nicol	Crépuscı	Auroralin	King Sap	Thérald	Malfred	Mabine	Érard	
Tienn		1	1	1	1	0	0	1	0	2	2	2	2	5	3	0	0	0	4	4	4	
Rakenn			1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Alyse				1	1	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	
Le Capitann					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Gauve						2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	
Méric							2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	
Rashkang								2	2	0	0	0	0	0	0	0	0	0	0	0	0	
Chéram									1	0	0	0	0	0	0	0	0	0	0	0	0	
Lescantelles										1	1	0	0	0	0	0	0	0	0	0	0	
La Rasinne											1	1	1	5	0	0	0	0	0	0	0	
Werfam												1	1	5	0	0	1	0	0	0	0	
Auguss													1	5	0	0	0	0	0	0	0	
Le Lamineur														5	0	0	0	0	0	0	0	
Nicol															0	0	0	0	0	0	0	
Crépusculine																1	1	1	0	0	0	
Auroraline																	1	1	0	0	0	
King Saphir																		1	0	0	0	
Thérald																			0	0	0	
Malfred																				4	4	
Mabine																					4	
Érard																						



#### **Application to a small dataset** Auguss Le Lamineur Lescantelles La Rasinne (21 characters) Le Capitanr Méric Rashkang Chéram ienn takenn Nerfam Gauve Alyse



ing Saphi

nérald Malfred

Part #2 Before the story begins Part #1 Part #3 Part #4

#### **Application to a small dataset** Auguss Le Lamineur L Alyse L Capitann Gauve Lescantelles La Rasinne Méric Rashkang (21 characters) Chéram Rakenn Nerfam



ing Saphi

nérald Malfred

Êrard

Part #2 Before the story begins Part #1 Part #3 Part #4

## Application to a small dataset (21 characters) (21 characters)

	Ę	enn	e	Capita	Ne	<u>.</u>	hkan	ram	cante	Rasin	fam	nss	amir	_	pusci	oralin	j Sap	rald	fred	oine	p
	Tien	Rak	Alys	Le	Gau	Mér	Ras	Ché	Les	LaF	Wer	Aug	Le L	Nico	Crél	Auro	Kinç	Thé	Mali	Mab	Éraı
Tienn		1	1	1	1	0	0	1	0	2	2	2	2	5	3	0	0	0	4	4	4
Rakenn			1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Alyse				1	1	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Le Capitann					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gauve						2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Méric							2	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Rashkang								2	2	0	0	0	0	0	0	0	0	0	0	0	0
Chéram									1	0	0	0	0	0	0	0	0	0	0	0	0
Lescantelles										1	1	0	0	0	0	0	0	0	0	0	0
La Rasinne											1	1	1	5	0	0	0	0	0	0	0
Werfam												1	1	5	0	0	1	0	0	0	0
Auguss													1	5	0	0	0	0	0	0	0
Le Lamineur														5	0	0	0	0	0	0	0
Nicol															0	0	0	0	0	0	0
Crépusculine											/					1	1	1	0	0	0
Auroraline										1	1						1	1	0	0	0
King Saphir										/								1	0	0	0
Thérald									1	/									0	0	0
Malfred									1											4	4
Mabine									/												4
Érard								17													









## Problem:

- The optimization of the order of 80 columns
- Factorial complexity
- ♦ 80! ≈ 10e118
- Previously published heuristic algorithm limited to 20-25 columns

## **AFB metaheuristic**

Artificial Feeding Birds (AFB) [Springer] Inspired by the behaviour of pigeons



- → Simple
- → Performant
- → Generic

An optimisation problem = A triplet ( coût(), vol(), marche() )

#### Same parameter functions as previously

Jondrette	Mme Burgon Child 1	Child 2	Mme Hucheloup Grantaire	Bahorel	Joly	Bossuet	Gavroche 🔴	Enjoiras Courfevrac	Jean Prouvaire	Combeferre	Feuilly	M Mabeuf	Motner Plutarcn I t Theodiile Gillenormand	Mile Vaubois	Mlle Gillenormand	Mme Pontmercy	George Pontmercy	Boulatruelle	Brujon Montnarnassa	Babet	Gueulemer	Claquesous	Eponine	Thènardier 🛑 Mme Thénardier	Anzelma	Magnon	M Gillenormand	Baroness of T	Marius Cosetta	Félix Tholomyès	Blacheville	Dahlia	Fameuil	Favourite	Listolier	Lepnine Toussaint	Old woman 2	Jacquin Labarre	Old woman 1	Sister Perpetue Sister Simulice	Eantine	Javert	Jean Valjean 🔴	M Bamatabois Brevet	Champmathieu	Chenildieu	Cocnepame	Petit Gervais	Isabeau	Mme de R	M Scaufflaire	Mother Innocent	Fauchelevent	Mlle Baptistine	Mme Magloire	Count Countess de Lô	Napoleon	M Myriei	Marquis de Champtercier	Cravatte	
																																																			G	B Sy re R	lu /a ?e	e n d	:   :   :   :   :	pa pa pa pa	urt urt urt urt	1 2 3 4 5	)		

Most new encounters occur in parts 1 and 3, and seems well-separated



Gavroche appears in part 3 but encounters many more characters in part 4





#### Another matrix

- Same novel and same characters
- Numbers in the matrix are the number of times the characters meet

## Different parameter functions

 $select: p \mapsto True \ if \ p \neq 0, False \ otherwise$  $aggregate: c_1, c_2, \dots, c_m \mapsto c_1 + c_2 + \dots + c_m$  $colorize: c \mapsto \begin{cases} blue & if \ \frac{c}{c_{max}} = 0\\ \dots \\ red & if \ \frac{c}{c_{max}} = 1 \end{cases}$ 







# Discussion

## Interesting for highly connected networks

Rainbow boxes are more compact that matrices

### We inferred groups of characters from pairwise relations

- A knows B, A knows C, B knows C
- ♦ => A, B, C know each others
- But is A aware that B knows C?

#### Perspectives

- Extend the method to non-symmetric matrices (directed graphs)
- Adapt the method to other overlapping set visualization techniques
  - Venn diagrams,...
- Apply the method to other domains
  - Social media (FOAF), bioinformatics (protein interaction matrices),...

# References

**[iV2016]** : Lamy JB, Berthelot H, Favre M. Rainbow boxes: a technique for visualizing overlapping sets and an application to the comparison of drugs properties. International Conference Information Visualisation (iV) 2016;:253-260

**[JVLC]** : Lamy JB, Berthelot H, Capron C, Favre M. Rainbow boxes: a new technique for overlapping set visualization and two applications in the biomedical domain. Journal of Visual Language and Computing 2017;43:71-82

**[Springer]** : Lamy JB. Artificial Feeding Birds (AFB): a new metaheuristic inspired by the behavior of pigeons. Advances in nature-inspired computing and applications 2018; under press

**[Leydesdorff]** : Leydesdorff L and Vaughan L. Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. Journal of the association for information science and technology 2006;57(12):1616-1628

**[Jänicke]** : Jänicke S, Franzini G, Cheema MF, and Scheuermann G. Visual text analysis in digital humanities. 2016;226-250

