A Netflix Tour of Data Science

Film Suggestion by Diffusion on Graphs

Avignon Edwige – Fourcade Pierre – Nguyen Kenneth

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Introduction

- Nowadays with the importance of cinema and the number of movies coming out each year, it is hard to even know what we want to watch !

- Streming platforms, such as Netflix, understood this and are investing on movies suggestion engines.

→ How, as students in data science, can we tackle the issue of film suggestion ?

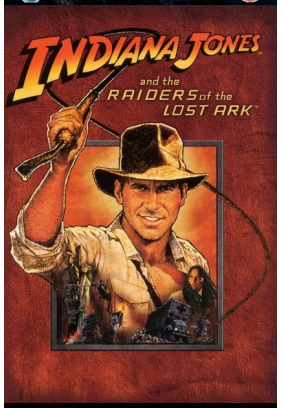
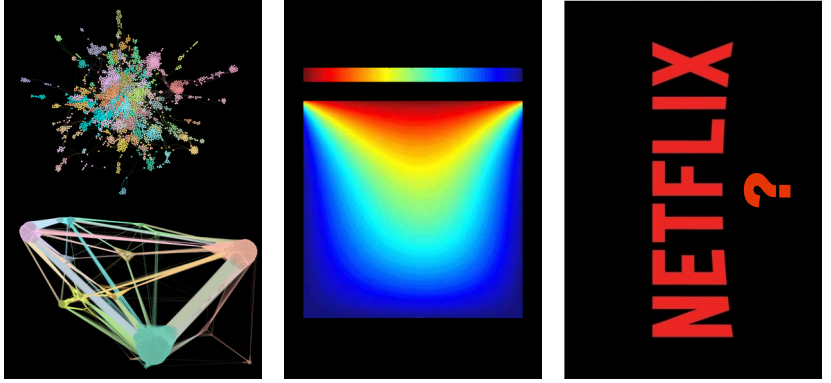
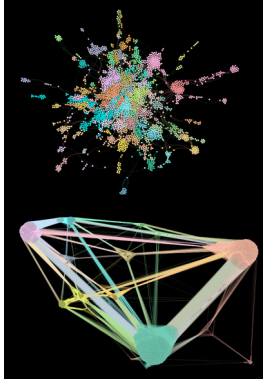


Table of contents



I. Why and how do we build our graphs ?

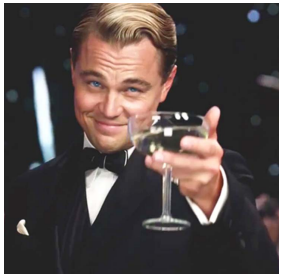
1. Purpose and features of the graphs



Idea :
A user who likes a film with particular characteristics is likely to appreciate other films with similar ones

→ Gather films with common characteristics

Genre | Main role | Cast | Crew

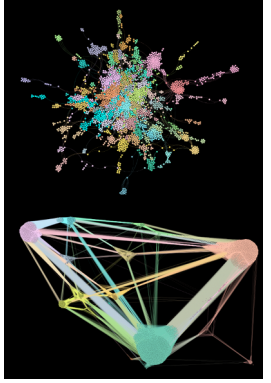


= important features that can influence the user's opinion about a film

→ Creation of a graph for each feature
nodes = films

→ Each graph has to represent similarities between movies in regards of the feature chosen.
We have to define weighted adjacency matrices

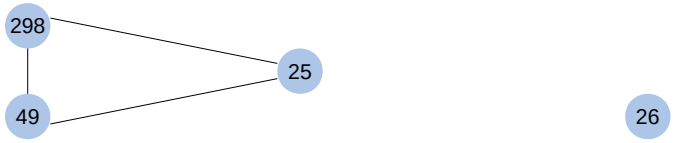
2. Defining similarities – Weighted adjacencies



To show similarities between movies, weights in the matrices have to be really well defined, weights have to make sense in a graph of movies.

Feature = Main role

	title	cast_name
25	Titanic	[Kate Winslet, Leonardo DiCaprio, Frances Fish...
26	Captain America: Civil War	[Chris Evans, Robert Downey Jr., Scarlett Joha...
298	The Wolf of Wall Street	[Leonardo DiCaprio, Jonah Hill, Margot Robbie,...
49	The Great Gatsby	[Leonardo DiCaprio, Tobey Maguire, Carey Mulli...

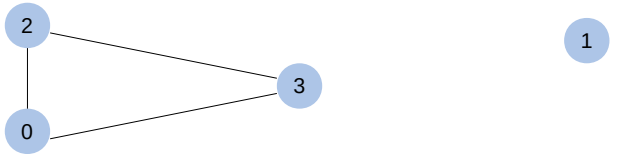


Feature = Cast

	title	cast_name
113	Harry Potter and the Order of the Phoenix	[Daniel Radcliffe, Rupert Grint, Emma Watson, ...
114	Harry Potter and the Goblet of Fire	[Daniel Radcliffe, Rupert Grint, Emma Watson, ...
115	Hancock	[Will Smith, Charlize Theron, Jason Bateman, J...
189	Noah	[Russell Crowe, Jennifer Connelly, Emma Watson

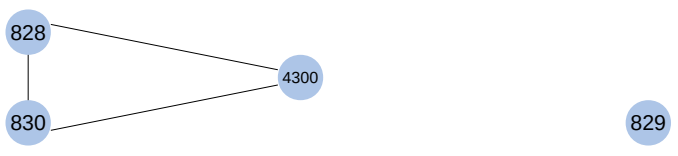
Feature = Genre

	title	genres_name
0	Avatar	[Action, Adventure, Fantasy, Science Fiction]
1	Pirates of the Caribbean: At World's End	[Adventure, Fantasy, Action]
2	Spectre	[Action, Adventure, Crime]
3	The Dark Knight Rises	[Action, Crime, Drama, Thriller]



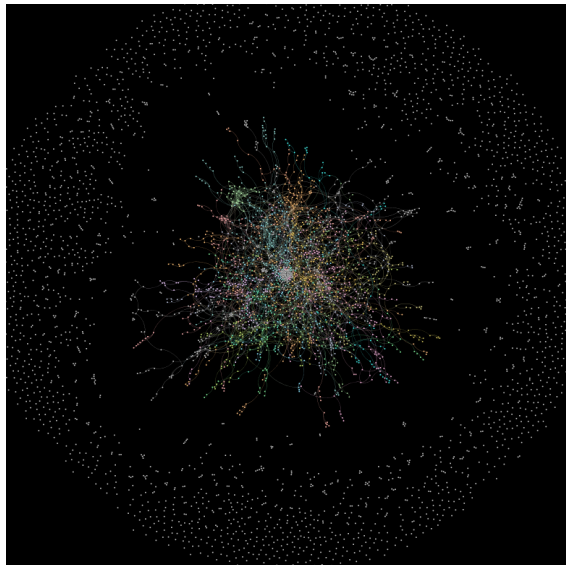
Feature = Crew

	title	crew_name	job
828	Kill Bill: Vol. 1	[Robert Rodriguez, Quentin Tarantino, Quentin ...	[Thanks, Director, Characters, Writer, Produce...
829	Bowfinger	[Brian Grazer, Bernard Williams, Bernard Willi...	[Producer, Executive Producer, Unit Production...
830	Kill Bill: Vol. 2	[Robert Rodriguez, Quentin Tarantino, Quentin ...	[Original Music Composer, Director, Screenplay...
4300	Reservoir Dogs	[Quentin Tarantino, Quentin Tarantino, Lawrenc...	[Director, Writer, Producer, Production Manage...

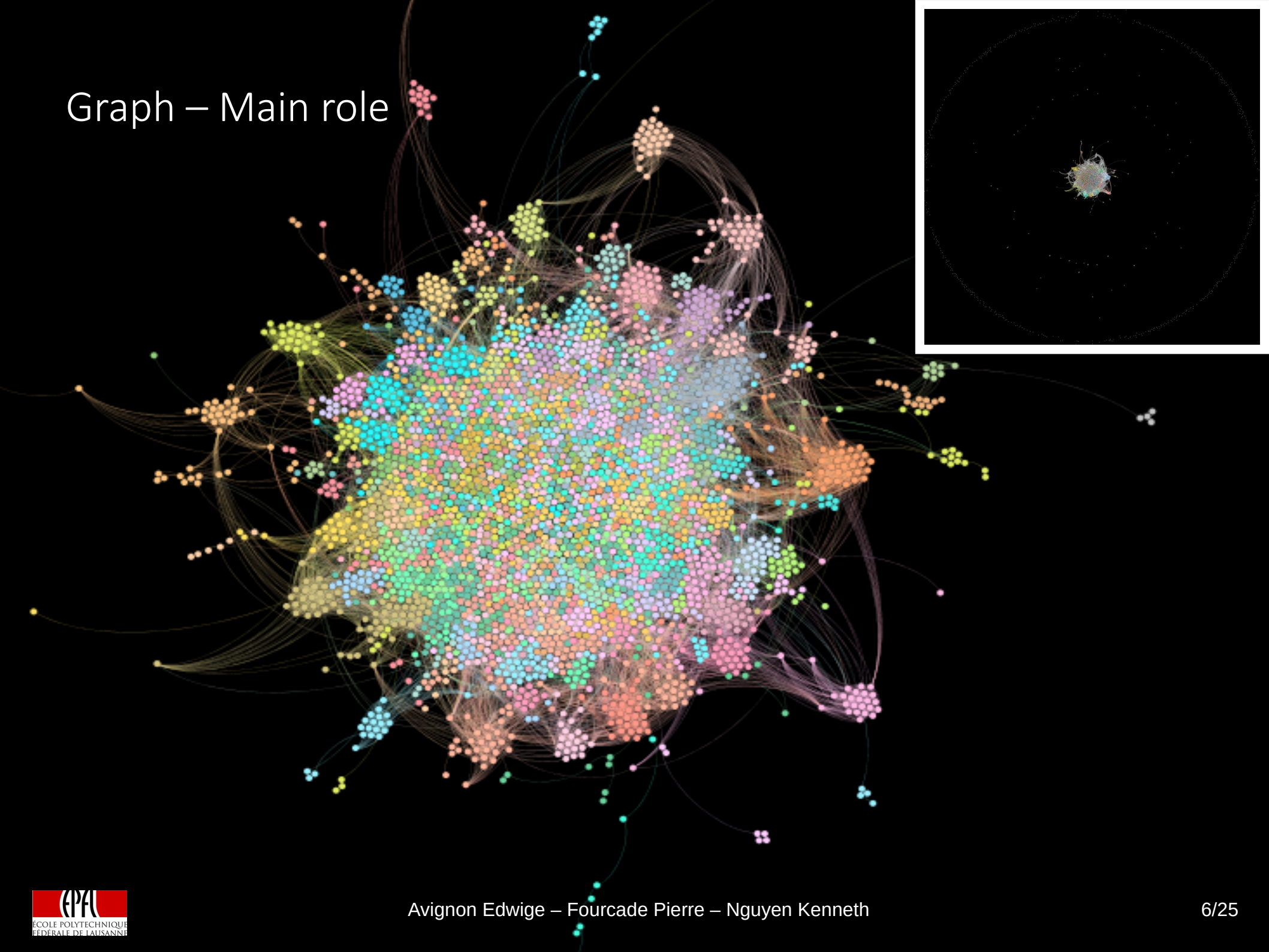


3. Results – Observations with Gephi

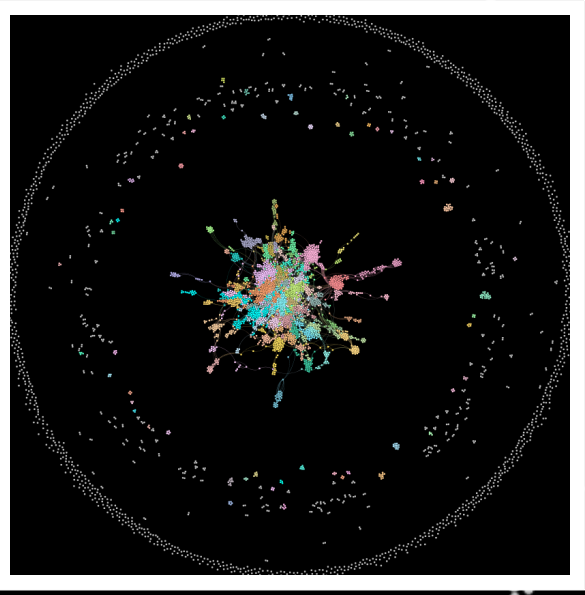
Graph- Cast



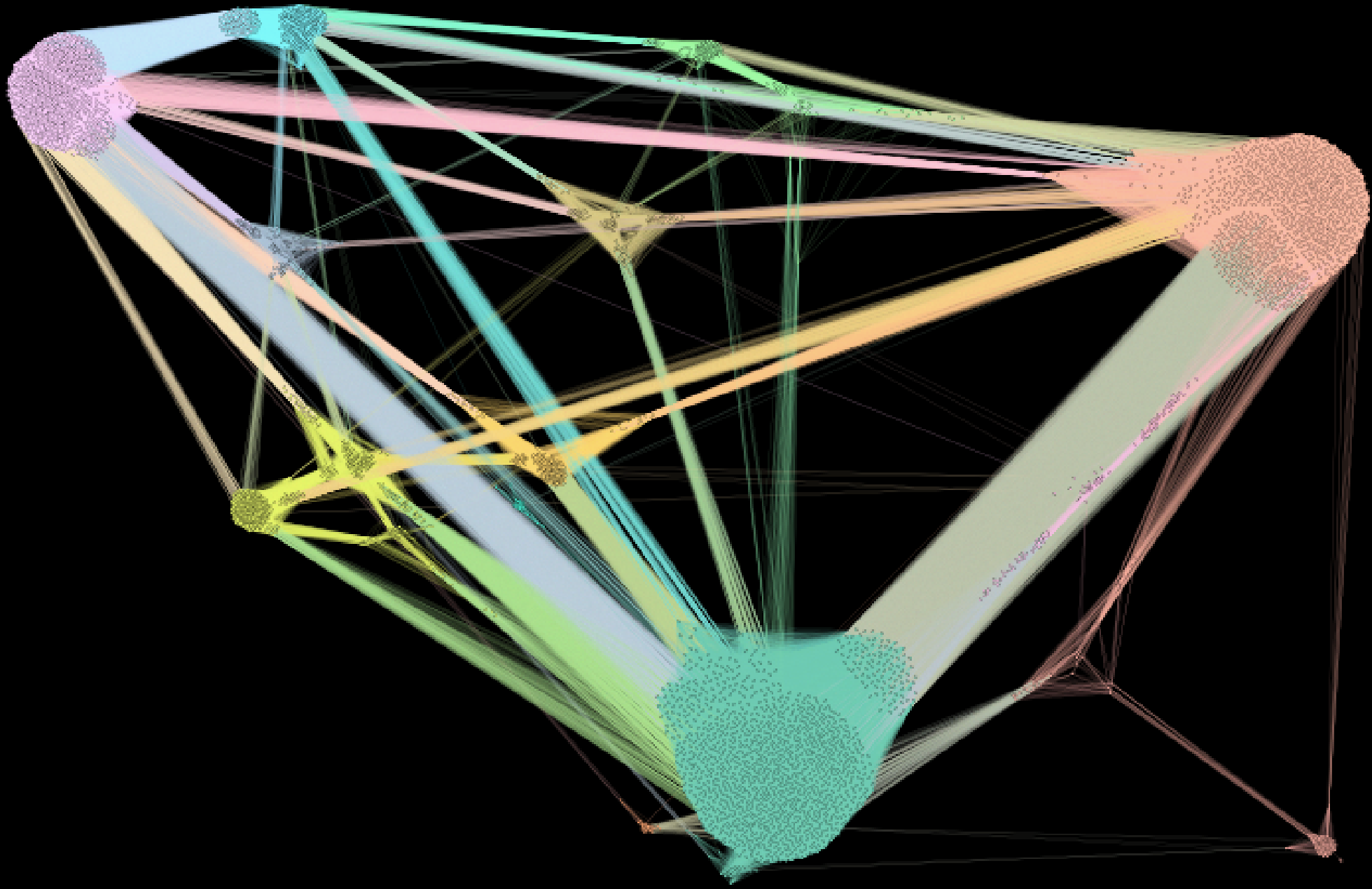
Graph – Main role



Graph- Crew

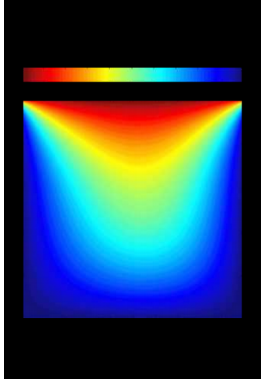


Graph – Genres



II. Adapted heat diffusion

1. Justification of the model



Idea :

Propose films to a user that he is likely to appreciate = films with similar features to films he has already watched and gave a good mark.

A solution :

Propagate the user's vote for film A to neighbours of film A → Heat diffusion

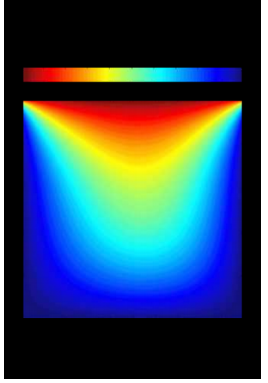
For our algorithm to propose relevant films but also diversified, the user would have to watch and rate a lot of movies.

A solution :

Associate the rates of our user (= **user's signal**) and the average rate of all the users (= **base signal**)

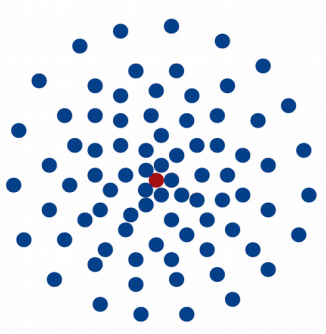
Without the base signal, the algorithm could suggest you a film with the same cast, same crew and same genre … which could be your biggest disappointment

2. Description of the model

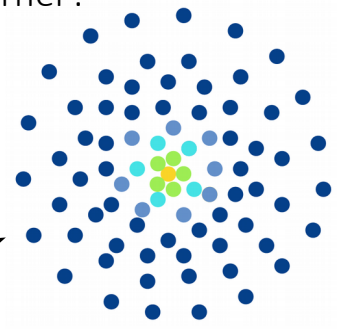


A modified heat diffusion
For example the user gave a 10, we diffuse this vote into the base signal

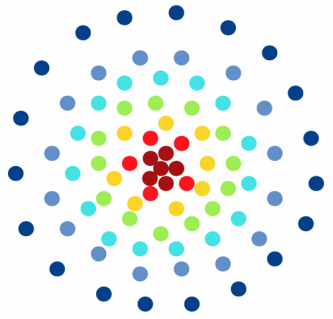
User's vote :



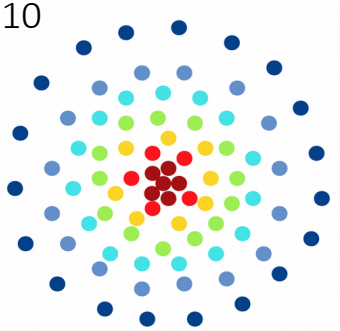
With diffusion by Heat kernel :



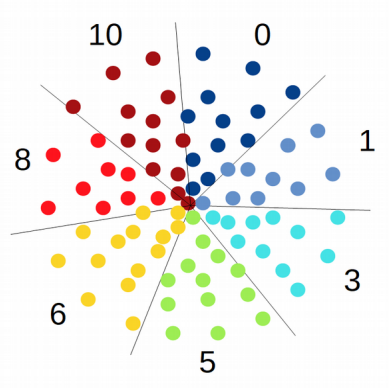
With diffusion with fixed vote of the user :



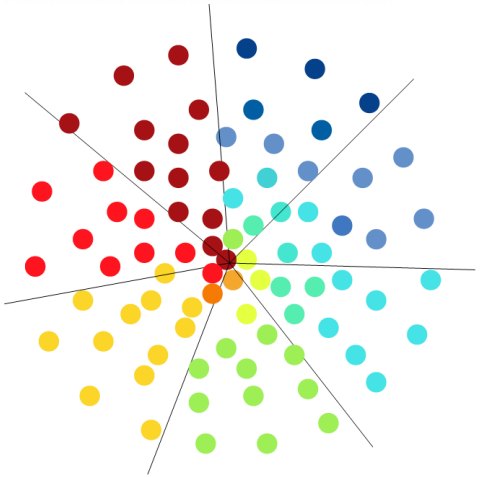
User's signal Normalized from 0 to 10

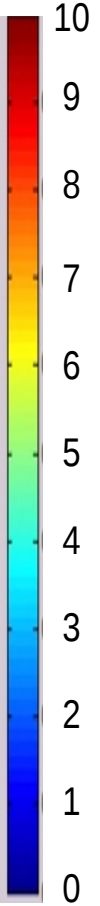


Base signal



- Mean of the 2 values
- But the user's signal cannot « cold » the signal because we placed a heat source





Same process with « cold source ». Graphs will change little by little as the user rates other films.

III. Tests and results

1. Test of each graph

a. Cast – Harry Potter

Test on the Harry Potter movies available in the dataset:

```
- Node number 8: Harry Potter and the Half-Blood Prince
- Node number 113: Harry Potter and the Order of the Phoenix
- Node number 114: Harry Potter and the Goblet of Fire
- Node number 191: Harry Potter and the Prisoner of Azkaban
- Node number 197: Harry Potter and the Philosopher's Stone
- Node number 276: Harry Potter and the Chamber of Secrets
```

If we give a 10 to the first movie of the saga, the suggestions obtained without base signal are :

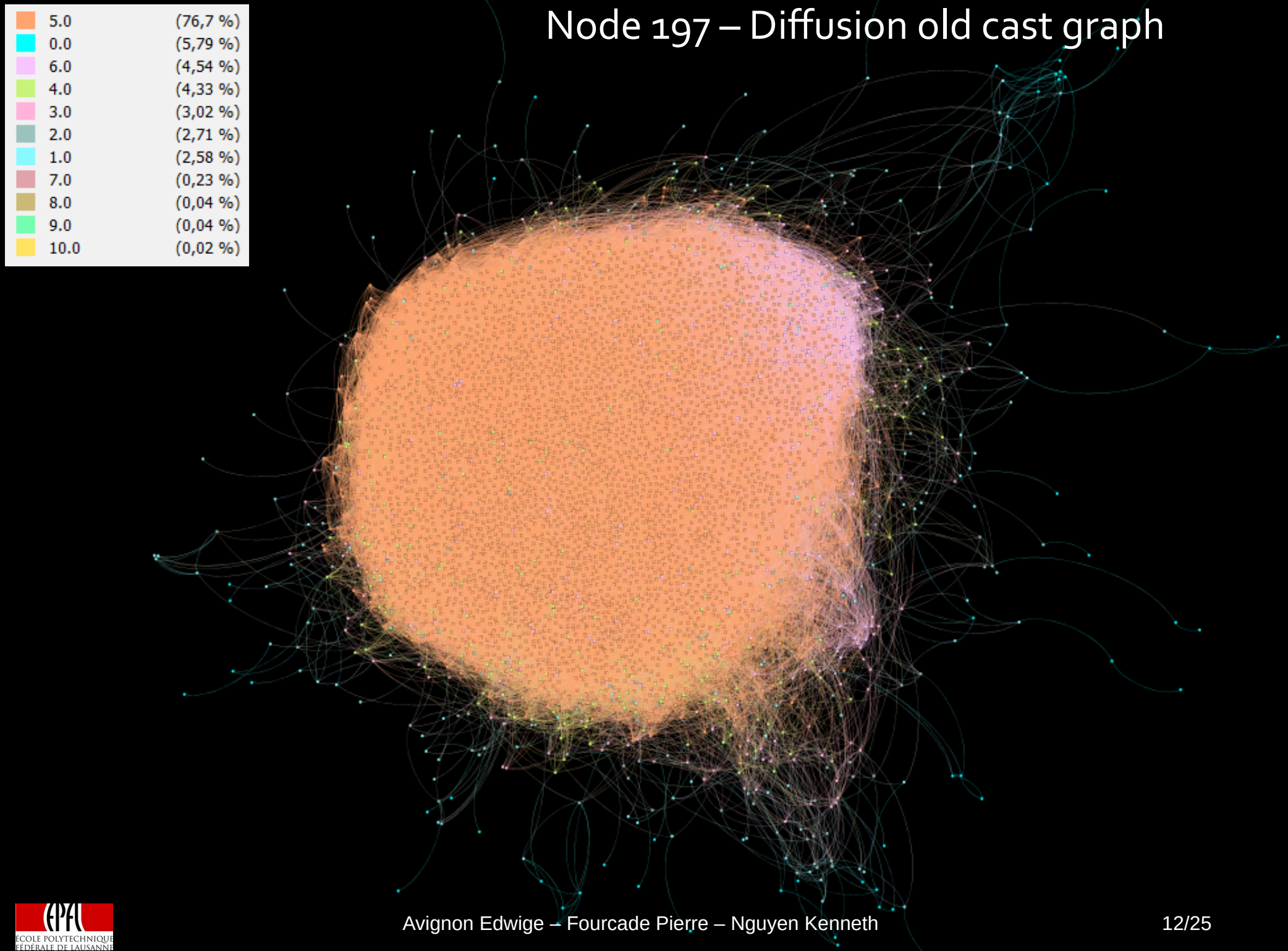
```
According to your likings you may want to take a look at:

- The Naked Gun 2½: The Smell of Fear
- The Apparition
- The History Boys
- The Boy
- Grabbers
- Quigley Down Under
- The Naked Gun 33⅓: The Final Insult
- The Devil's Double
- The Man from Snowy River
- Son of God
- Bloody Sunday
- Containment
- Rabbit-Proof Fence
- Recess: School's Out
- The Descent
- Airplane!
- Lesbian Vampire Killers
- The Woman in Black
- Risen
- Womb
```

- We are not able to find the rest of the films from the saga. Even though cast well defined !
- We have a problem with the first graph that we have built.

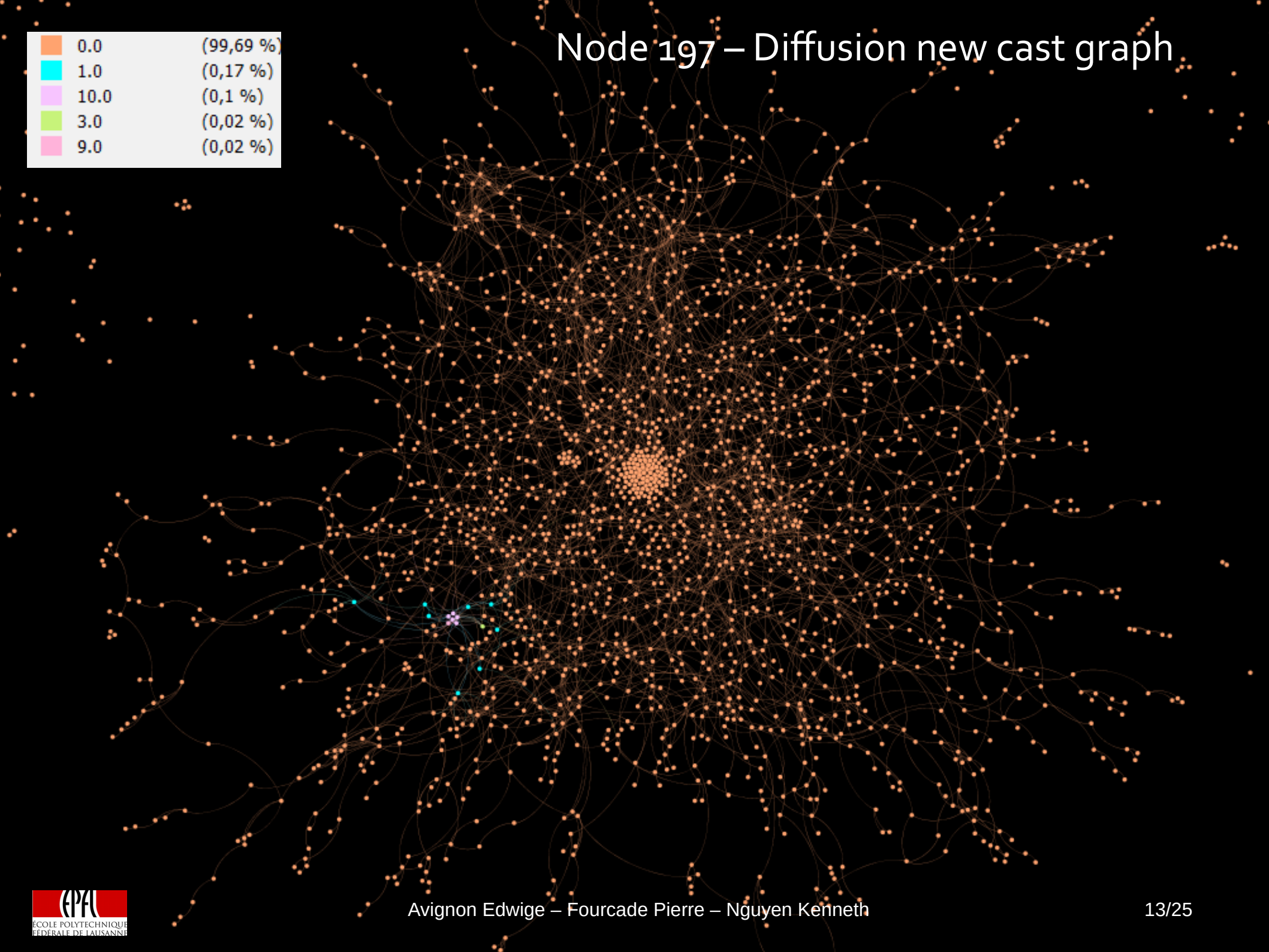
  Let's take a look at the graph.

Note: The graph slide 5 corresponds to the updated version made from this test.

Node 197 – Diffusion old cast graph



Node 197 – Diffusion new cast graph



III. Tests and results

1. Test of each graph

a. Cast – Harry Potter

- Here are the results by considering the updated version of the cast graph.

➡ The results here are much better !

- We understand here the importance of defining properly the weights and by extension the similarity.

```
According to your likings you may want to take a look at:

- Harry Potter and the Chamber of Secrets
- Harry Potter and the Prisoner of Azkaban
- Harry Potter and the Order of the Phoenix
- Harry Potter and the Half-Blood Prince
- Harry Potter and the Goblet of Fire
- The Tale of Despereaux
- Michael Collins
- Quartet
- The King's Speech
- Gosford Park
- A Room with a View
- The Book of Eli
- Alice Through the Looking Glass
- My Week with Marilyn
- The Curse of the Were-Rabbit
- Richard III
- Nanny McPhee and the Big Bang
- The Borrowers
- Regression
- Brave
```

III. Tests and results

1. Test of each graph

b. Main role – Leonardo DiCaprio

- Test on movies with Leonardo DiCaprio as the first role:

```
- Node number 49: The Great Gatsby
- Node number 96: Inception
- Node number 176: The Revenant
- Node number 250: The Aviator
- Node number 297: Blood Diamond
- Node number 298: The Wolf of Wall Street
- Node number 316: Gangs of New York
- Node number 351: The Departed
- Node number 439: Shutter Island
- Node number 622: Body of Lies
- Node number 883: Catch Me If You Can
- Node number 961: The Beach
- Node number 1081: Revolutionary Road
- Node number 1380: The Man in the Iron Mask
- Node number 1409: J. Edgar
- Node number 2661: Romeo + Juliet
```

- We give a 10 to « The Great Gatsby ».

- We obtained the following results:

```
According to your likings you may want to take a look at:

- The Beach
- The Man in the Iron Mask
- Blood Diamond
- Inception
- Revolutionary Road
- Shutter Island
- The Wolf of Wall Street
- The Revenant
- Gangs of New York
- The Aviator
- J. Edgar
- Romeo + Juliet
- Body of Lies
- The Departed
- Catch Me If You Can
- The Quick and the Dead
- Celebrity
- The Mighty
- What's Eating Gilbert Grape
- Basic Instinct 2
```

➡ Relevant results !

III. Tests and results

1. Test of each graph

c. Genre – Documentary

- Test on movies mentionned as documentaries for the first genre.
  Here is an extract of some:

```
- Node number 489: Oceans
- Node number 1730: Winged Migration
- Node number 2191: Capitalism: A Love Story
- Node number 2729: Justin Bieber: Never Say Never
- Node number 2837: Katy Perry: Part of Me
- Node number 3036: One Direction: This Is Us
- Node number 3168: Sicko
- Node number 3190: Glee: The Concert Movie
- Node number 3354: Michael Jordan to the Max
- Node number 3446: Fahrenheit 9/11
- Node number 3471: Dolphins and Whales: Tribes of the Ocean
- Node number 3503: Lake of Fire
- Node number 3557: Jackass: The Movie
- Node number 3626: Sea Rex 3D: Journey to a Prehistoric World
- Node number 3646: Standard Operating Procedure
- Node number 3751: What the #$*! Do We (K)now!?
- Node number 3768: The Real Cancun
- Node number 3855: Time to Choose
- Node number 3861: March of the Penguins
```

- We give a 10 to « Oceans ».

- We obtained the following results:

```
According to your likings you may want to take a look at:

- Mad Hot Ballroom
- March of the Penguins
- An Inconvenient Truth
- I Want Your Money
- Give Me Shelter
- Sea Rex 3D: Journey to a Prehistoric World
- Ayurveda: Art of Being
- 51 Birch Street
- Butterfly Girl
- Riding Giants
- A LEGO Brickumentary
- Indie Game: The Movie
- The Harvest (La Cosecha)
- The Hadza:  Last of the First
- Dogtown and Z-Boys
- Trekkies
- Born to Fly: Elizabeth Streb vs. Gravity
- What the #$*! Do We (K)now!?
- Sex With Strangers
- My Date with Drew
```

➡ Relevant results !

III. Tests and results

1. Test of each graph

d. Crew – Quentin Tarantino

- Test on the movies directed by Quentin Tarantino:
- We give a 10 to « Inglorious Basterds ».

```
- Node number 287: Django Unchained
- Node number 571: Inglourious Basterds
- Node number 684: The Hateful Eight
- Node number 828: Kill Bill: Vol. 1
- Node number 830: Kill Bill: Vol. 2
- Node number 880: Grindhouse
- Node number 2822: Jackie Brown
- Node number 3232: Pulp Fiction
- Node number 3766: Four Rooms
- Node number 4300: Reservoir Dogs
```

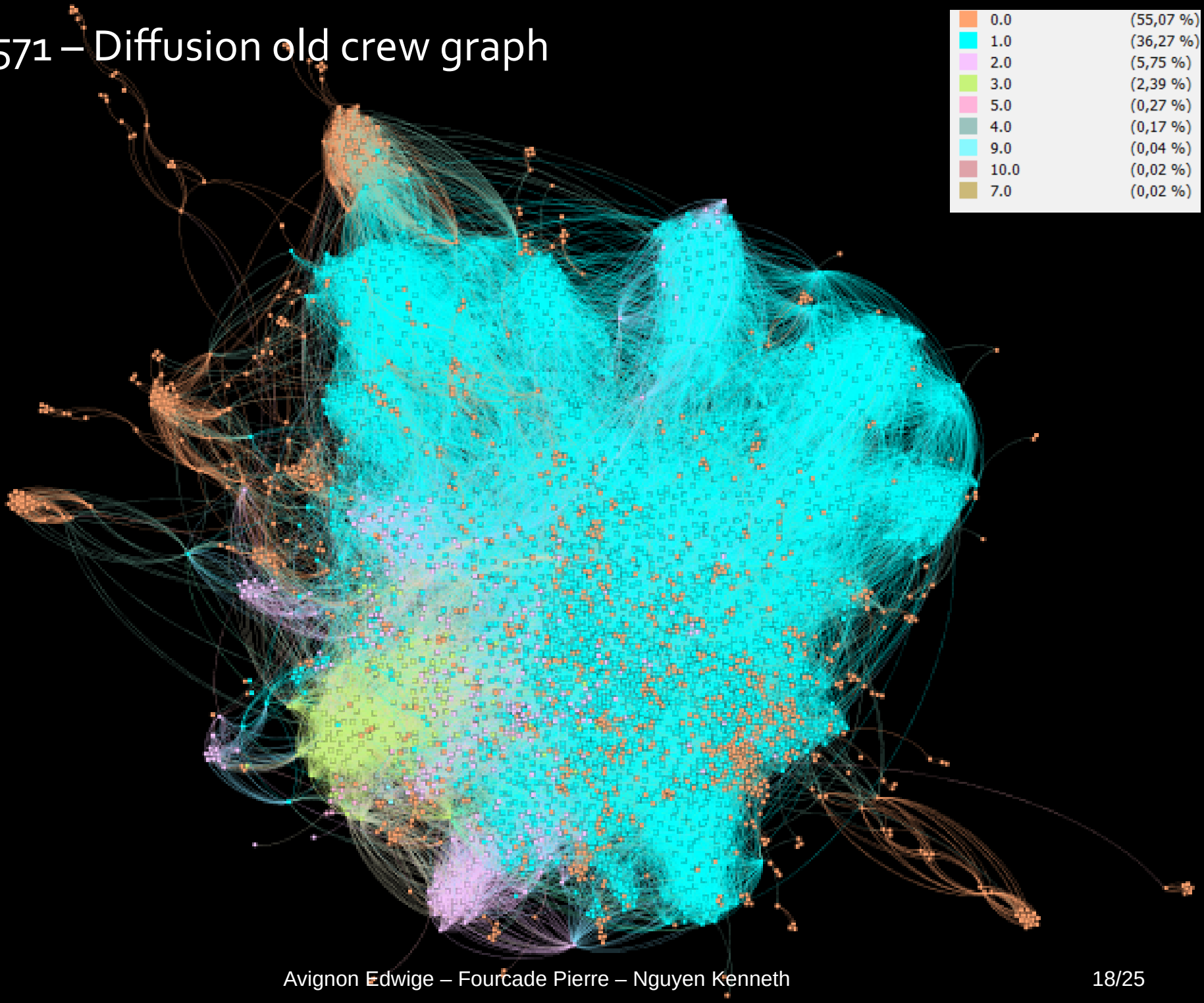
```
According to your likings you may want to take a look at:

- Jackie Brown
- Reservoir Dogs
- Killing Zoe
- Four Rooms
- Scary Movie 5
- Garden State
- Machete Kills
- Chasing Amy
- Once Upon a Time in Mexico
- Yoga Hosers
- Clerks II
- El Mariachi
- Desperado
- Hostel
- Hostel: Part II
- Mallrats
- Cop Out
- Machete
- Scary Movie 4
- Shorts
```

- We are able to find three of his movies.
- Seems to be too little in comparison to the other graphs.

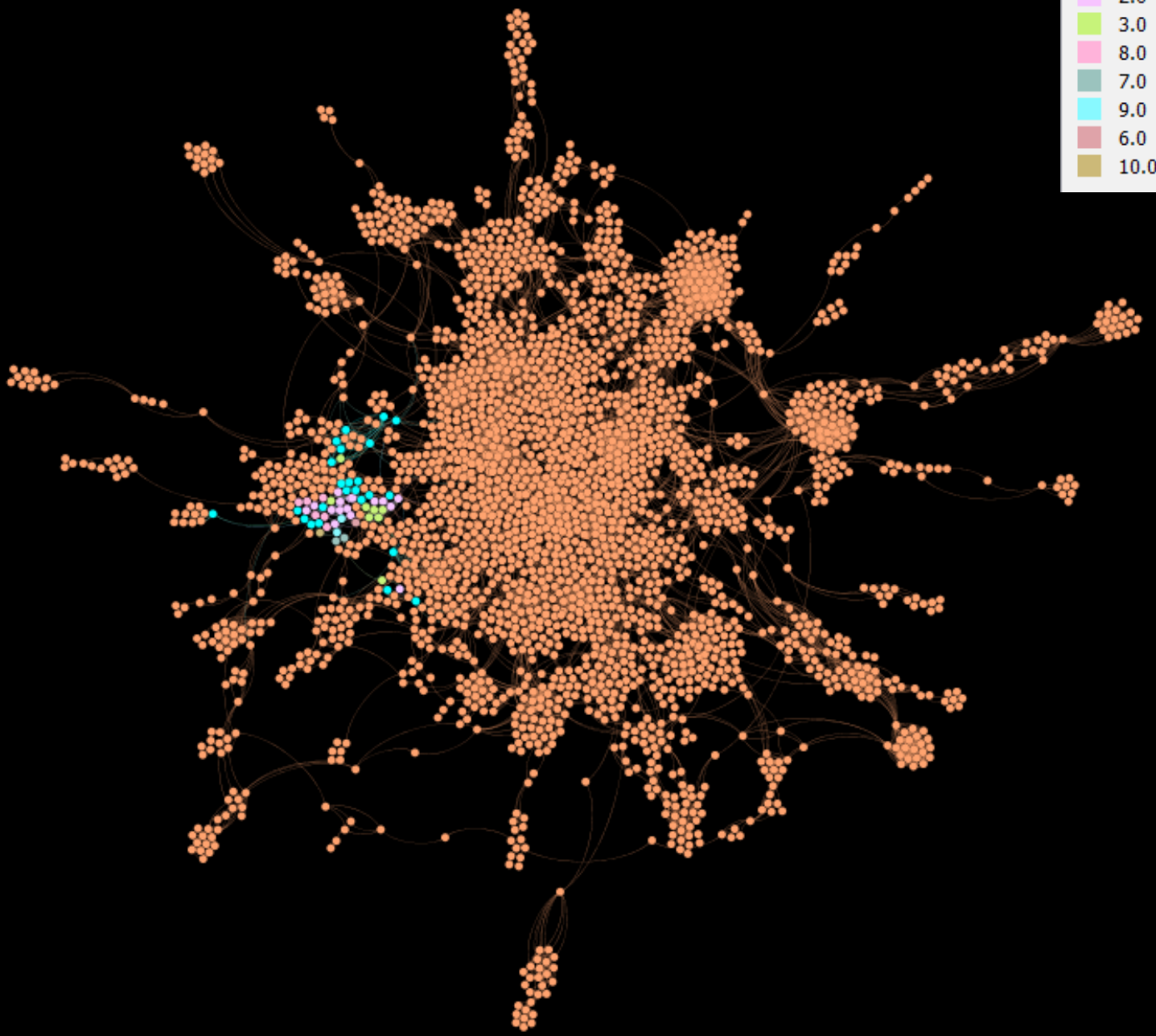
➡ Let's take a look at the graph.

Note: The graph slide 7 corresponds to the updated version made from this test.

Node 571 – Diffusion old crew graph



Node 571 – Diffusion new crew graph



III. Tests and results

1. Test of each graph

d. Crew – Quentin Tarantino

- Here are the results by considering the updated version of the crew graph.
- The results are better but not perfect still in comparison to the other graphs.
- To not see « Pulp Fiction » is normal. The movie is among the suggestions from the other users.

```
According to your likings you may want to take a look at:

- Reservoir Dogs
- Jackie Brown
- Kill Bill: Vol. 2
- Kill Bill: Vol. 1
- Django Unchained
- Ghost Dog: The Way of the Samurai
- The Man with the Iron Fists
- Grindhouse
- Hobo with a Shotgun
- Four Rooms
- The Hateful Eight
- Desperado
- Scott Pilgrim vs. the World
- Soul Plane
- El Mariachi
- Shorts
- Spy Kids: All the Time in the World
- Once Upon a Time in Mexico
- Machete
- Machete Kills
```

III. Tests and results

2. Results of the full method

We use the same tests as before but now with our full method.

Harry Potter 1 liked by the user :

```
According to your likings you may want to take a look at:

- Princess Mononoke
- Harry Potter and the Chamber of Secrets
- Harry Potter and the Prisoner of Azkaban
- The Lord of the Rings: The Return of the King
- Interstellar
- The Lord of the Rings: The Two Towers
- The Lord of the Rings: The Fellowship of the Ring
- Harry Potter and the Goblet of Fire
- Back to the Future
- Once Upon a Time in America
- The Dark Knight
- The Green Mile
- Psycho
- Seven Samurai
- Howl's Moving Castle
- Anne of Green Gables
- The Empire Strikes Back
- GoodFellas
- The Visual Bible: The Gospel of John
- American History X
```

➡ Relevant results !

Three Leonardo DiCaprio liked by the user :

```
According to your likings you may want to take a look at:

- Shutter Island
- American History X
- Anne of Green Gables
- The Visual Bible: The Gospel of John
- Psycho
- GoodFellas
- Once Upon a Time in America
- The Dark Knight
- Seven Samurai
- Guten Tag, Ramón
- Dead Poets Society
- Room
- The Usual Suspects
- City of God
- The Wolf of Wall Street
- The Departed
- Modern Times
- The Work and the Glory II: American Zion
- It's a Wonderful Life
- Samsara
```

➡ Relevant results !

III. Tests and results

2. Results of the full method

10 documentaries liked by the user :

```
According to your likings you may want to take a look at:

- Roger & Me
- The Dark Knight
- Anne of Green Gables
- GoodFellas
- Howl's Moving Castle
- The Visual Bible: The Gospel of John
- The Green Mile
- The Empire Strikes Back
- Psycho
- American History X
- Once Upon a Time in America
- Seven Samurai
- Princess Mononoke
- Bowling for Columbine
- Interstellar
- Inception
- Modern Times
- City of God
- Room
- Dead Poets Society
```

2 movies directed by Tarantino liked by the user:

```
According to your likings you may want to take a look at:

- American History X
- Anne of Green Gables
- The Visual Bible: The Gospel of John
- Psycho
- The Dark Knight
- Once Upon a Time in America
- GoodFellas
- Dead Poets Society
- Guten Tag, Ramón
- Room
- City of God
- The Usual Suspects
- Modern Times
- Reservoir Dogs
- The Work and the Glory II: American Zion
- It's a Wonderful Life
- Samsara
- The Prestige
- Apocalypse Now
- The Pianist
```

➡ Relevant results more difficult to obtain !

III. Tests and results

3. Test with cold sources

- We test here the effect of cold sources on the graph.
- For that we create of false signal with the movies with Leonardo DiCaprio seen previously. Then we dislike (give a 0) to three of those movies:

```
According to your likings you may want to take a look at:

- The Departed
- J. Edgar
- Body of Lies
- The Great Gatsby
- The Man in the Iron Mask
- The Wolf of Wall Street
- Shutter Island
- The Beach
- The Aviator
- Revolutionary Road
- Inception
- Gangs of New York
- Catch Me If You Can
- The Revenant
- Romeo + Juliet
- Blood Diamond
- The Empire Strikes Back
- Psycho
- Princess Mononoke
- The Dark Knight
```

False signal with DiCaprio movies.

```
According to your likings you may want to take a look at:

- Howl's Moving Castle
- The Shining
- The Good, the Bad and the Ugly
- Once Upon a Time in the West
- The Green Mile
- Karachi se Lahore
- Iraq for Sale: The War Profiteers
- Lake of Fire
- Call + Response
- Rise of the Entrepreneur: The Search for a Better Way
- One Direction: This Is Us
- Tupac: Resurrection
- Dolphins and Whales: Tribes of the Ocean
- The Lion King
- Se7en
- Blade Runner
- The Last Waltz
- Alien
- Eternal Sunshine of the Spotless Mind
- 2001: A Space Odyssey
```

➡ No more movies with DiCaprio !

III. Tests and results

4. Test – Influence of the base signal

- We test here if the base signal has the intended influence.
- For that we take the Indiana Jones movies. In the dataset with can find the 2nd, the 3rd and … the 4th. We like the 2nd and the 3rd and we hope to not get the 4th as a suggestion:

Results with the base signal:

```
According to your likings you may want to take a look at:

- The Empire Strikes Back
- Princess Mononoke
- Raiders of the Lost Ark
- Star Wars
- Jurassic Park
- The Lord of the Rings: The Return of the King
- Interstellar
- Return of the Jedi
- The Lord of the Rings: The Fellowship of the Ring
- The Lord of the Rings: The Two Towers
- Back to the Future
- Saving Private Ryan
- Big Hero 6
- Howl's Moving Castle
- Once Upon a Time in America
- American History X
- The Green Mile
- Psycho
- GoodFellas
- The Dark Knight
```

Pleasant surprise !

Results without the base signal:

```
According to your likings you may want to take a look at:

- Raiders of the Lost Ark
- Indiana Jones and the Kingdom of the Crystal Skull
- Return of the Jedi
- Star Wars
- The Empire Strikes Back
- Jurassic Park
- War of the Worlds
- The BFG
- Hook
- The Adventures of Tintin
- The Lost World: Jurassic Park
- The Fugitive
- Babe: Pig in the City
- Star Wars: Episode I - The Phantom Menace
- Star Wars: Episode II - Attack of the Clones
- Mad Max 2: The Road Warrior
- Mad Max
- 1941
- Seven Years in Tibet
- Force 10 from Navarone
```

Conclusion

With the different tests and improvements the results seem to be relevant.

Ideas for improvement:

- build more graphs to take into account more features (release date, budget …)
- make more analysis of the influence of the diffusion parameter τ of the heat-kernel
- investigate on other approaches for the diffusion
- add a collaborative filtering algortihm, used by Netflix

Avignon Edwige – Fourcade Pierre – Nguyen Kenneth

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE