



Sweet A sociolinguistics of Twitter

<https://sosweet.inria.fr>

E. Fleury, ENS de Lyon / Inria

<https://team.inria.fr/dante/>

CLARIN-PLUS workshop — Creation and Use of Social Media Resources, May 2017

Outline

- Objectives of SoSweet
- Data collections
- Tools
- Future Challenges

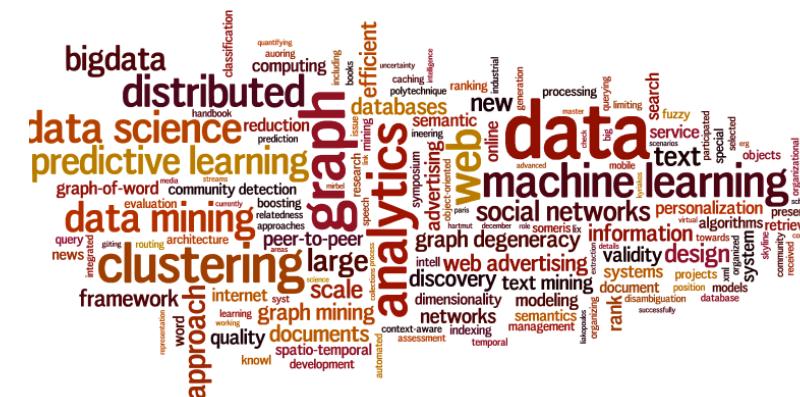
Objectives #1

- Provide a detailed understanding of the dynamic links between:
 - individuals,
 - social structure,
 - and language variation and change
- through the study of:
 - synchronic variation
 - diachronic evolution
 - of the variety of French language observed on



Objectives #2

- Develop
 - interdisciplinary,
 - computational and data driven approaches



- to handle the enormous amount of collected digital data specific to social media

Twitter as a lab

to study how social structure shapes
linguistic variability and change

Language on Twitter shows

- High variability
- High innovation rates

Twitter provides large amount of

- Linguistics data
- Social data

Not forgetting bias

- Technological bias
- Communicative bias

Linguistic Data

tweets

Social Data

Followers network

Corpus linguistics

Sociolinguistics

Natural Language
Processing

Network science

Synchrony

Linguistic variation /
Social structure

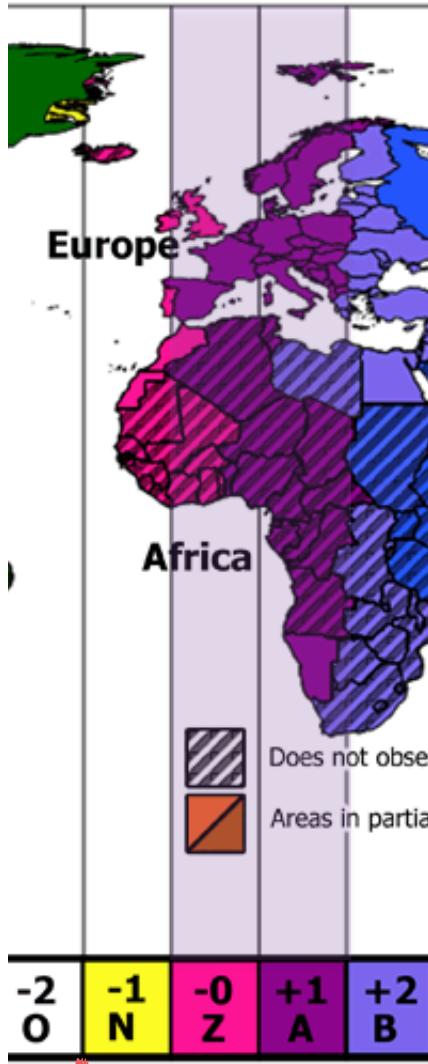
Diachrony

Language change /
Social structure

Outline

- Objectives of SoSweet
- Data collections
- Tools
- Future Challenges

Data collection: tweets



I tweet in French

I leave in
GMT or
GMT+1

Target 500 million tweets

- 200 millions
- annotated with parts of speech
- Gnip/Datasift/JSON

Processed / TAG (CC Tagset)

June 2014

Now

Dec 2019

Data collection: networks



Data collection: socio-economic data



I answer fun quiz and some demographic questions

<http://sosweet.ish-lyon.cnrs.fr>



Sauriez-vous deviner qui se cache derrière chaque tweet ?



#francophonie



#argot



#people

Outline

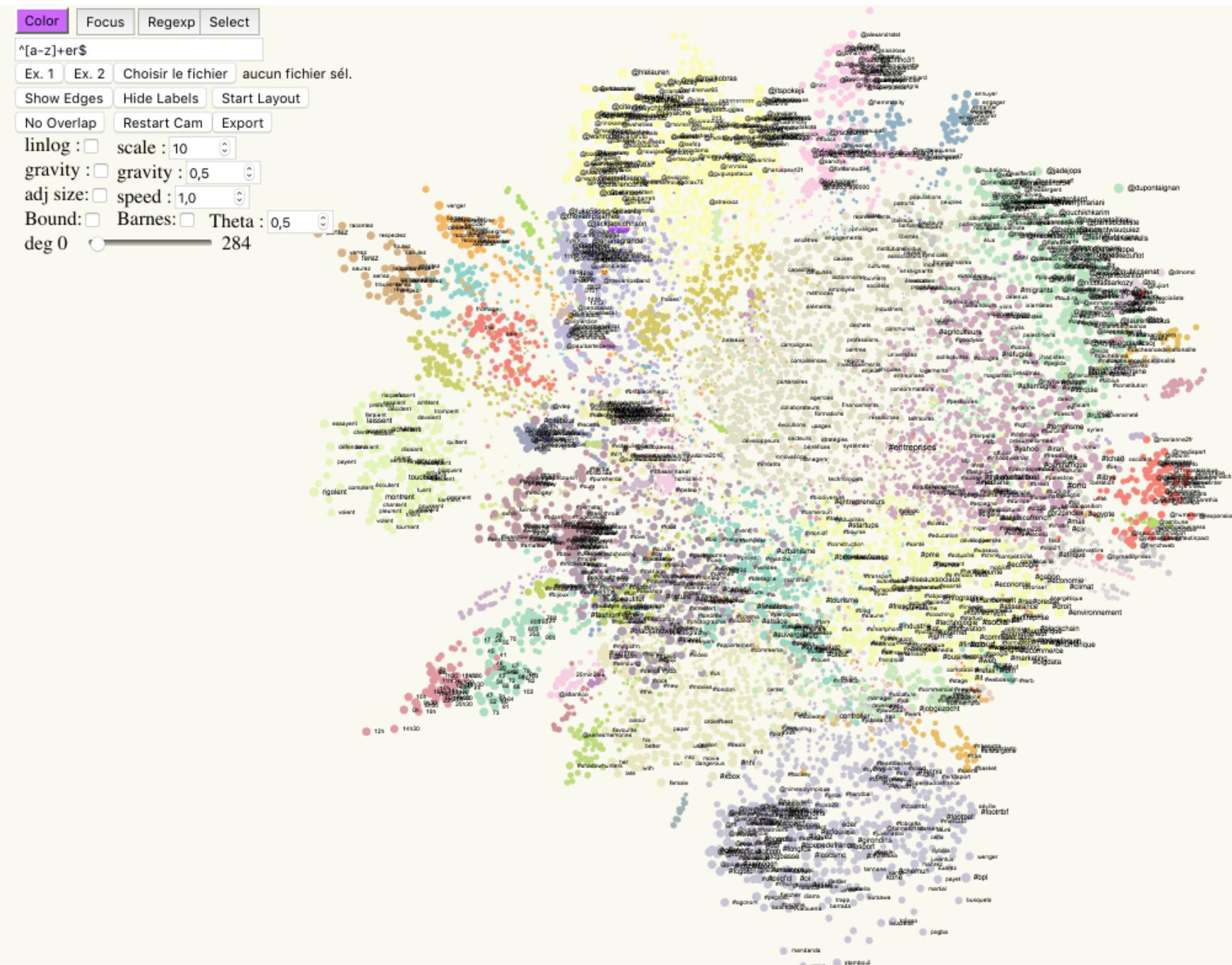
- Objectives of SoSweet
- Data collections
- Tools
- Future Challenges

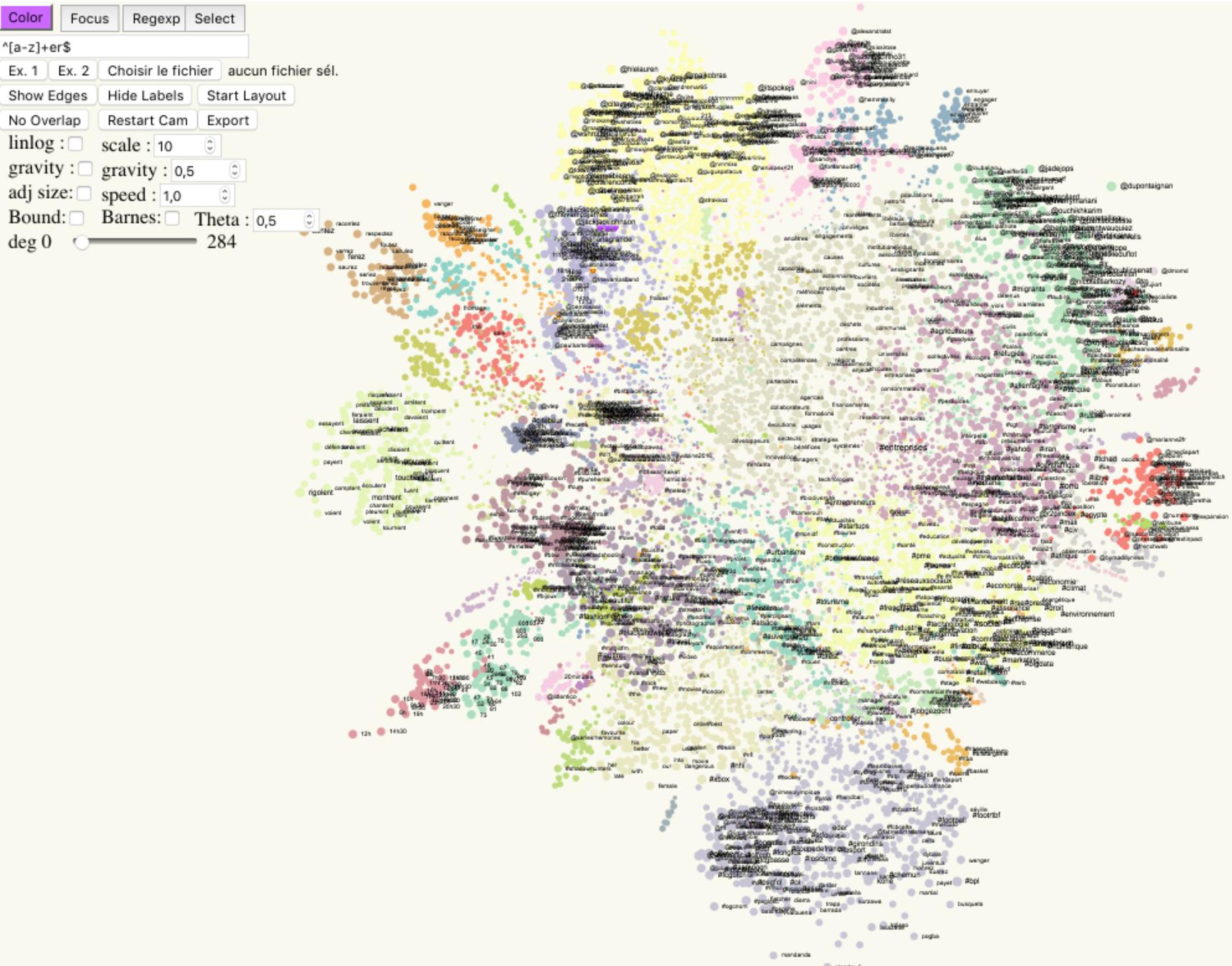
ALMAnaCH Linguistic workbench

- MELT:Part Of Speech tagger
 - Coupling annotated corpus / morphosyntactic lexicon
- FRMG – French Meta Grammar
 - Wide coverage abstract grammatical description for French
- Word Embeddings with Glove/FRMG
 - Global Vectors for Word Representation
 - syntax link / not only co-ocurrence
- <https://gforge.inria.fr/projects/lingwb/>

Word2Graph embedding

- Use Glove or word2vect
 - embedding of the vocabulary to a « space »
- Build a k-closest proximity graph on
 - Two word close in the space are link
 - Non symmetric relation
- Run community detection on the proximity graph
 - <https://gitlab.inria.fr>





Color Focus Regexp Select

`^[a-z]+er$`

Focus Regexp Select

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sél.

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

linlog : scale : 10

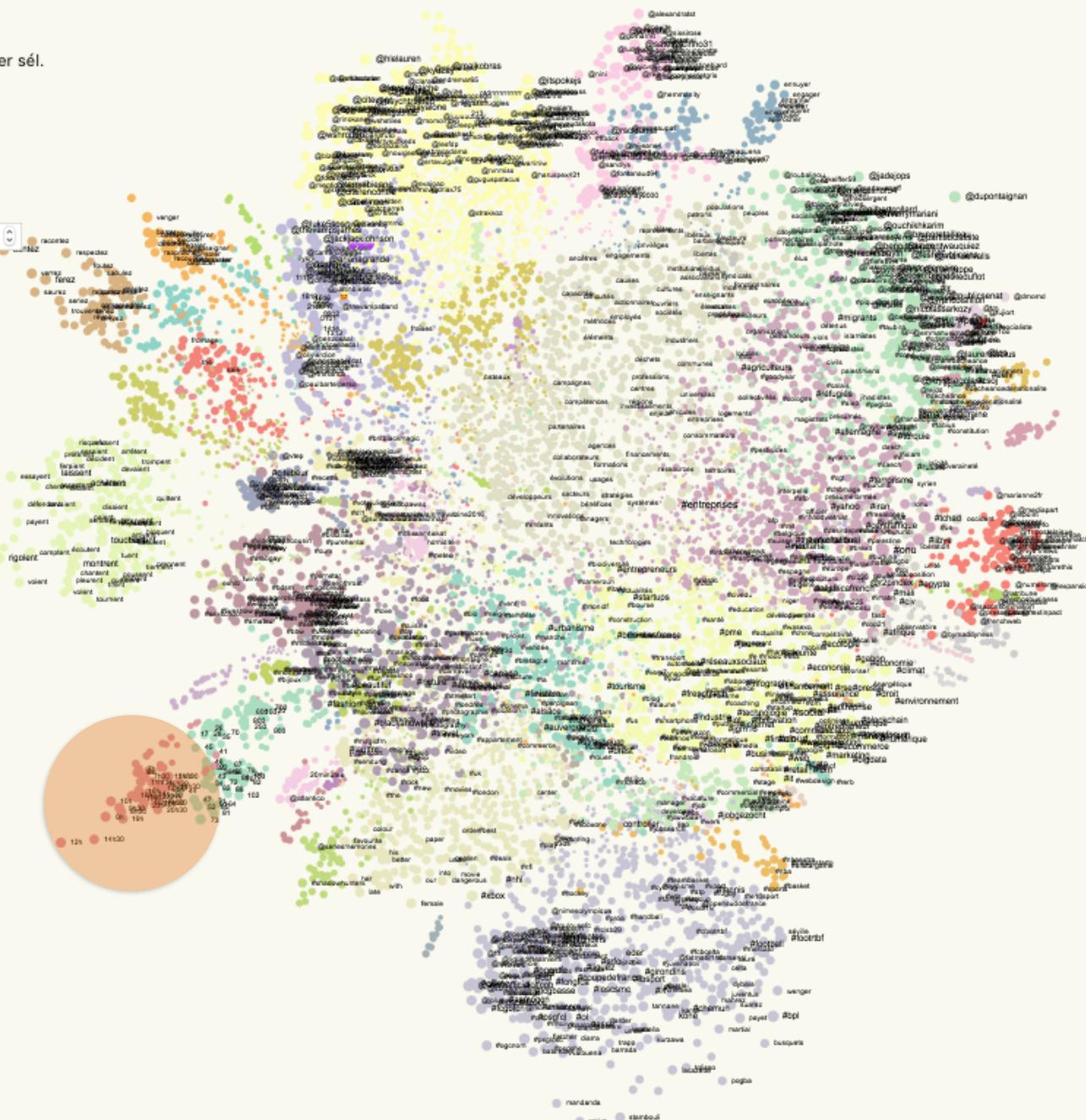
gravity : gravity : 0,5

adj size: speed : 10

Bound: Barnes: T

deg 0 284

deg 0 284



Color Focus Regex Select

`^[a-z]+er$`

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sél.

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

linlog : scale : 10

gravity : gravity : 0,5

adj size: speed : 1,0

Bound: Barnes: Theta : 0,5

deg 0

284

3

44

47

52

17

30min

1h

1h30

2h30

2h

3h

4h

5h

6h30

7h

8h30

9h30

10h

9h

19h

8h30

18h

17h30

10h30

16h

15h

23h

14h

13h

13h30

21h

16h30

12h30

22h30

00h4h30

5h30

15h30

15h30

15h30

21h30

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

30min

1h

1h30

2h30

2h

3

44

47

52

17

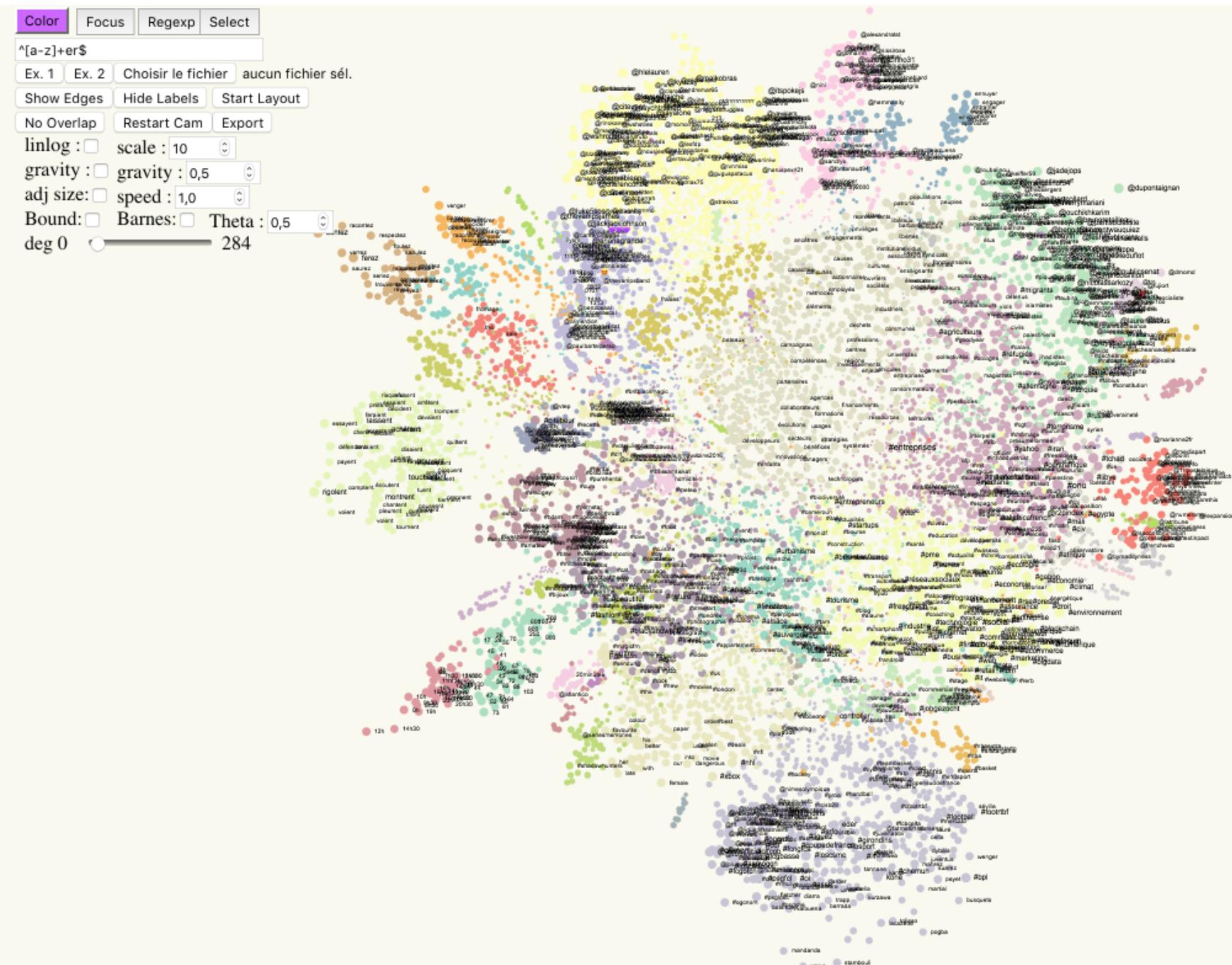
30min

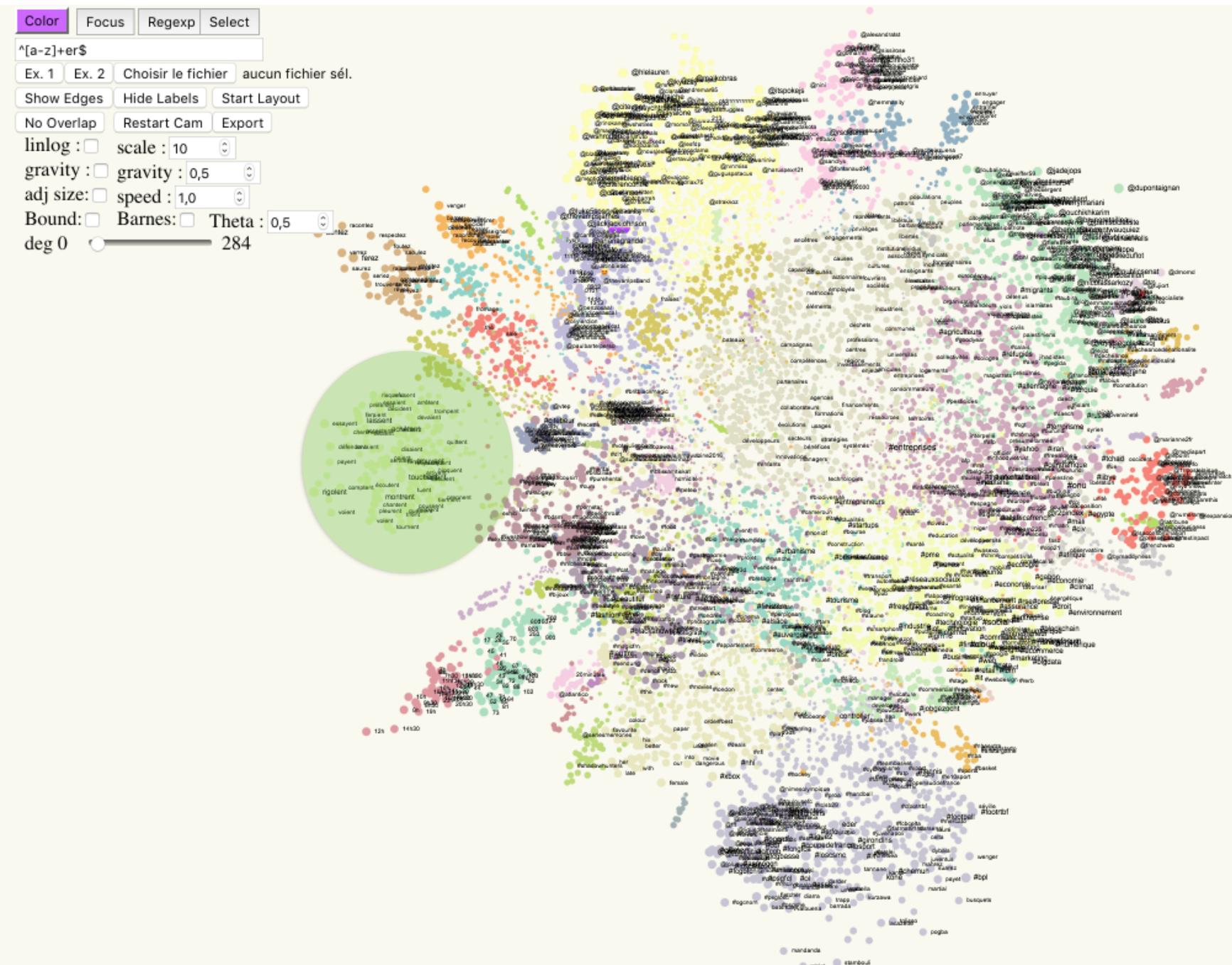
1h

1h30

2h30

2h





Color Focus Regex Select

^a-z]+er\$

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sél.

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

linlog : scale : 10

gravity : gravity : 0,5

adj size: speed : 1,0

Bound: Barnes: Theta : 0,5

deg 0 284

Chier aucun fichier sél.

Labels Start Layout

Start Cam Export

scale : 10

gravity : 0,5

speed : 1,0

Barnes: Theta : 0,5

284

essayent

osent

cherchent

défendent

savaient

forcent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

préfèrent

voulaient

feraient

laissent

découvrent

acceptent

osent

cherchent

osent

comptent

risquent

essaient

Color Focus Regexp Select

`^[a-z]+er$`

Focus Regexp Select

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sélec

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

linlog : scale : 10

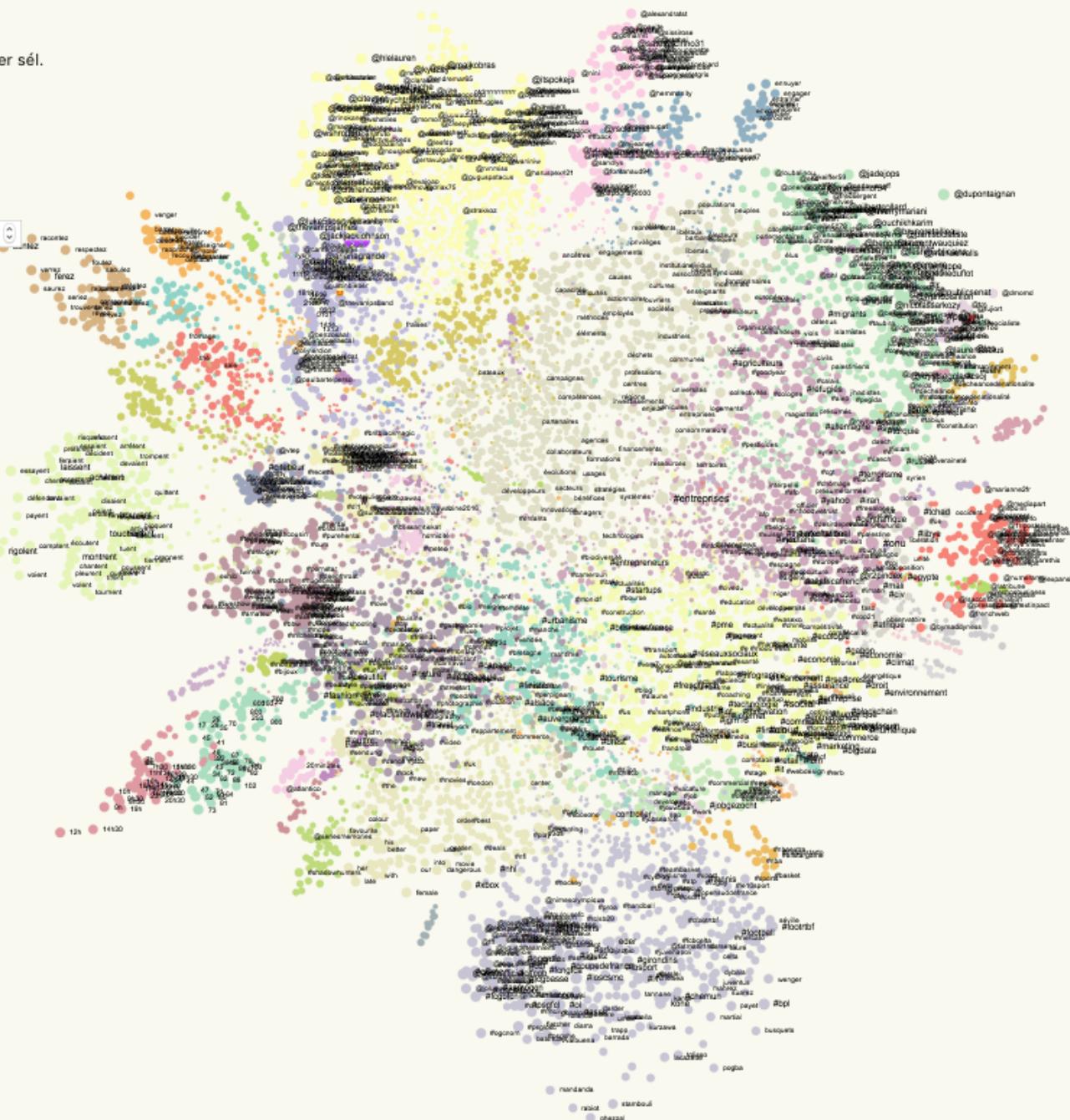
gravity : gravity : 0.5

adi size: speed : 10

adj size: speed : 1,0
Bound: Barnes: T

Bound: Barnes: Theta
doc 9 284

deg 0 284



Color Focus Regexp Select

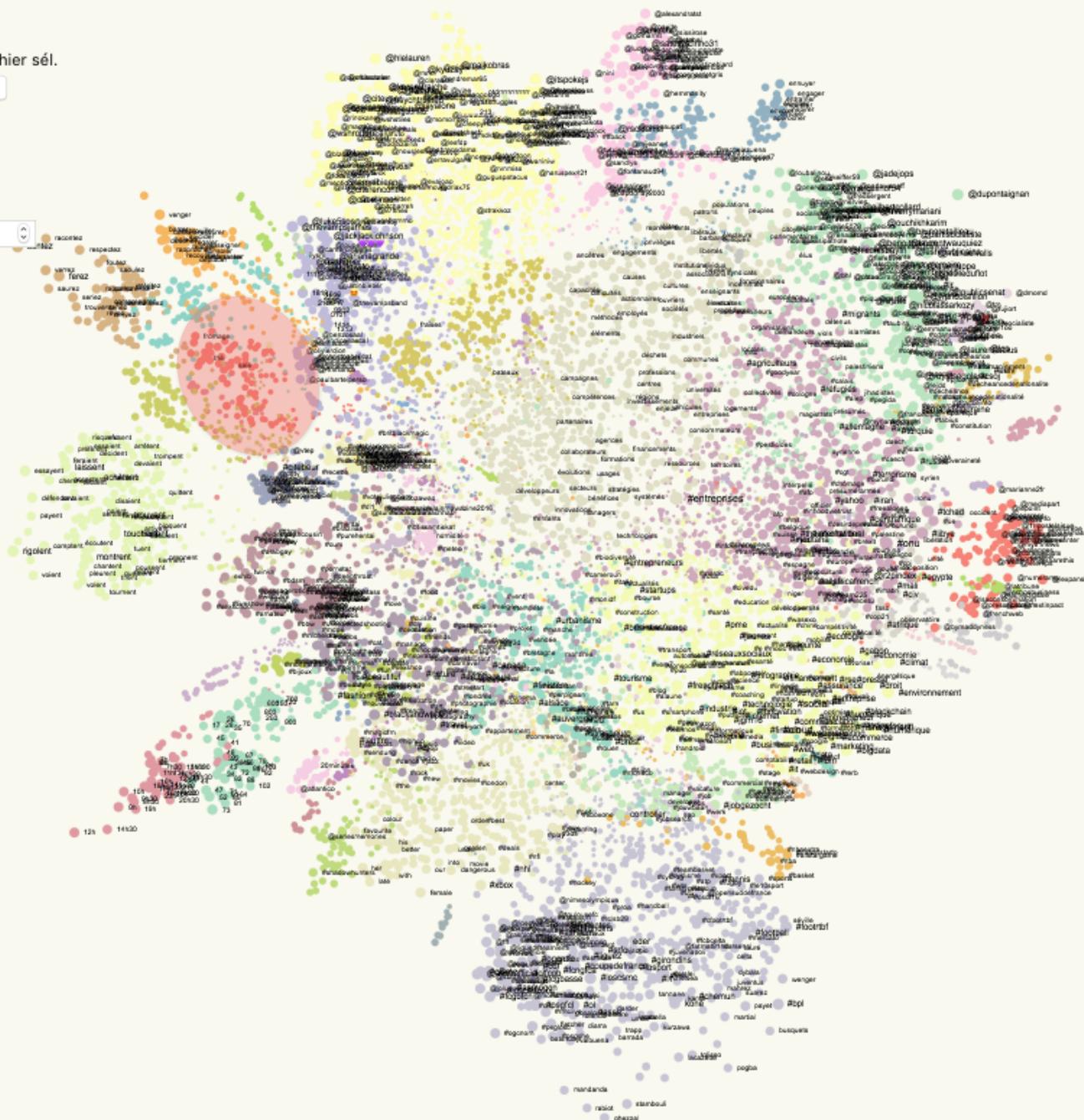
`^[a-z]+er$`

Focus Regexp Select

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sélec

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export



Color Focus Regexp Select

$^*[a-z]^+er\$$

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sé

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

linlog : scale : 10

gravity : a gravity : 0.5

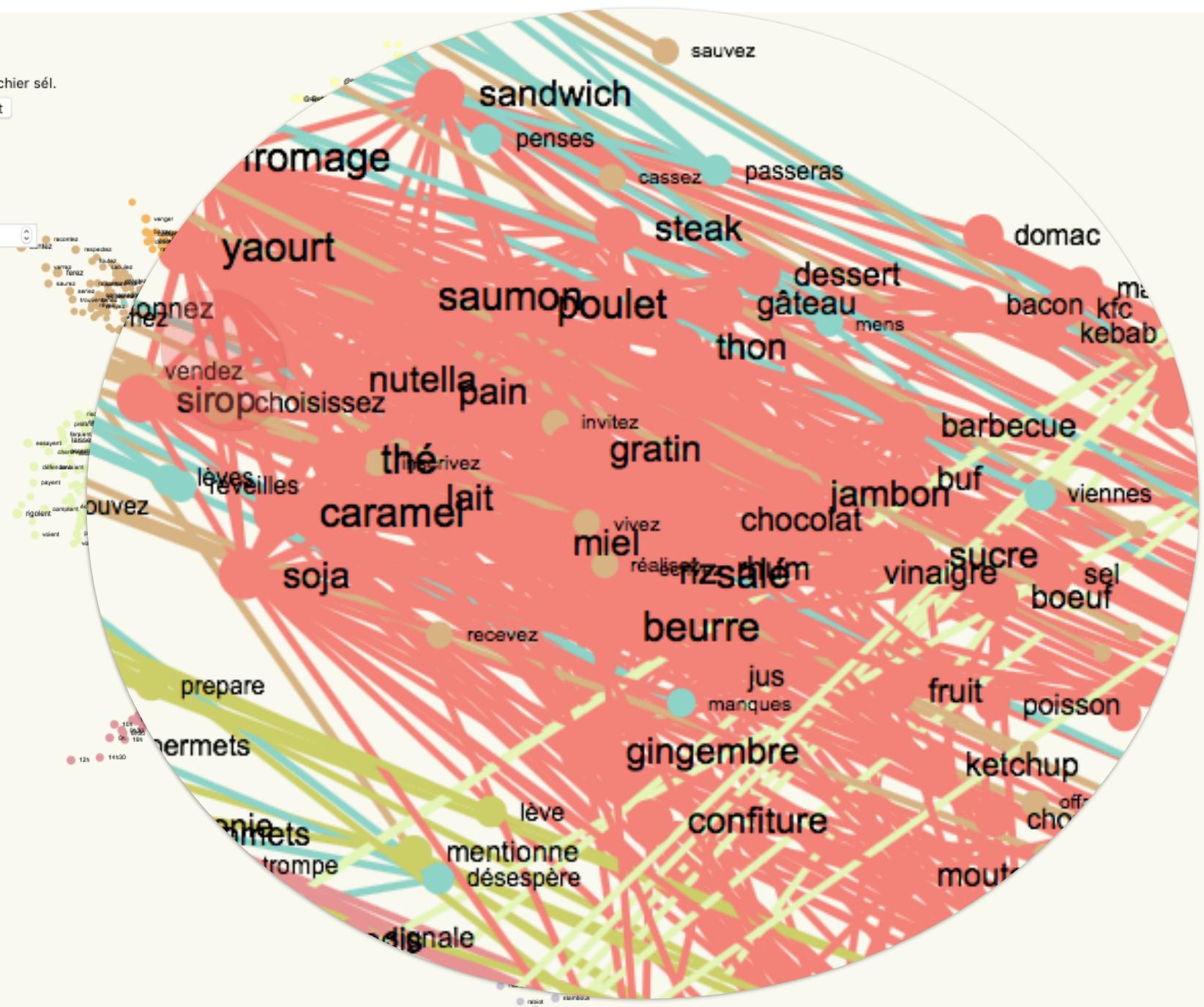
gravity : gravity : 0,5

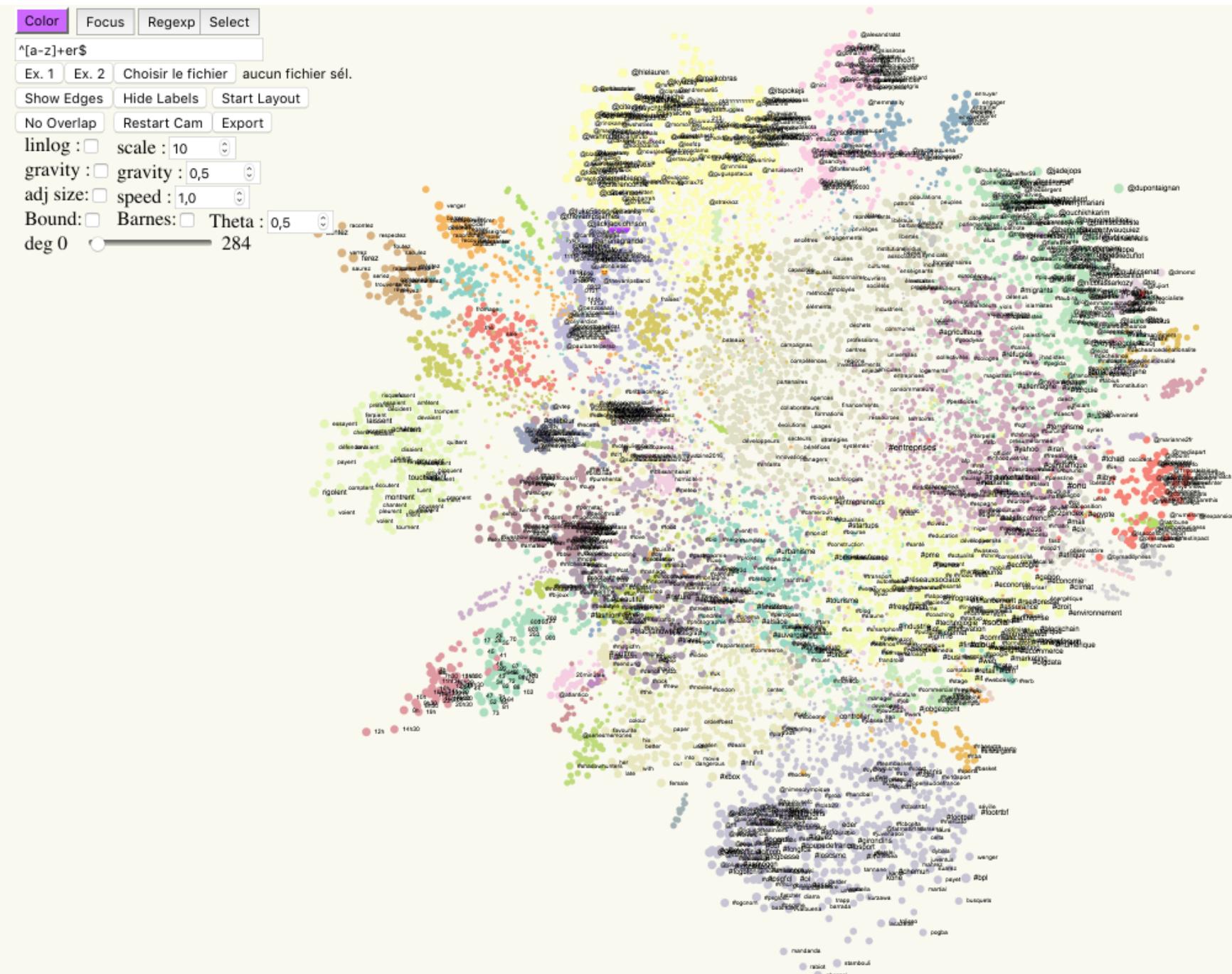
adj size: speed : 1,0

Bound: Barnes: Theta:

deg 0 ————— 284

and *the* *other* *two* *are* *not* *so* *far* *as* *they* *are* *concerned*





Color Focus Regex Select

^a-z]+er\$

Ex. 1 Ex. 2 Choisir le fichier aucun fichier sél.

Show Edges Hide Labels Start Layout

No Overlap Restart Cam Export

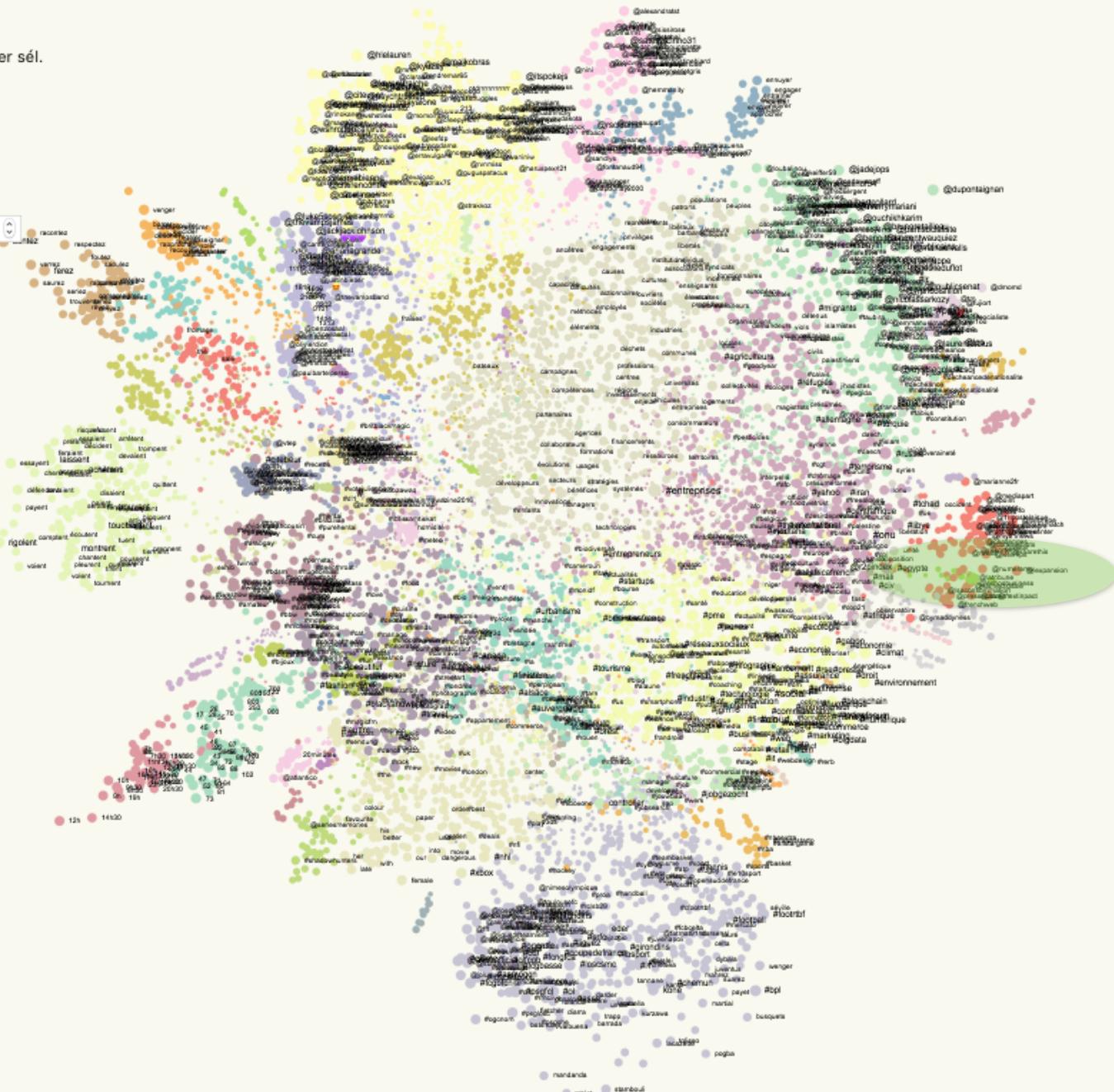
linlog : scale : 10

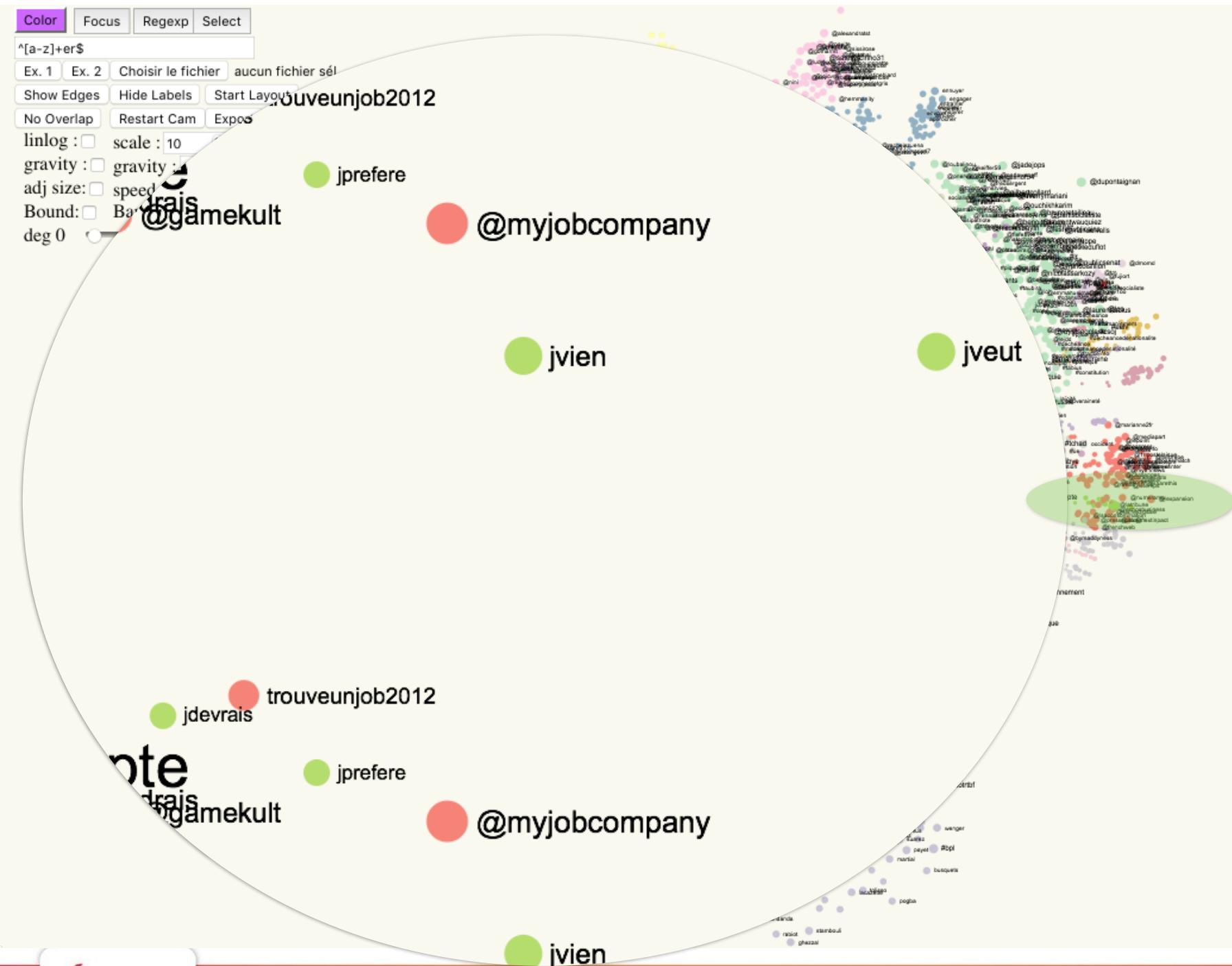
gravity : gravity : 0,5

adj size: speed : 1,0

Bound: Barnes: Theta : 0,5

deg 0





Outline

- Objectives of SoSweet
- Data collections
- Tools
- Future Challenges

Open SoSweet to the community

- Ethical issues
 - Operational Legal and Ethical Risk Assessment Committee (COERLE)
- Licensing issues / Twitter restriction
 - ☒ No full tweets distribution
 - ☑ Distributing IDs is allowed
- Aggregation
- Sampling
- On line tools for queries.



Thanks

ICAR

Matthieu Quignard
Sandra Teston-Bonnard
Clément Thibert
Nathalie Rossy-Gensane
Daniel Valéro

Alpage

Marie Candito
Eric de la Clergerie
Benoît Crabbé
Benoît Sagot
Djamé Seddah

Lidilem

Jean-Pierre Chevrot
Aurélie Nardy
Julie Peuvergne

Dante

Eric Fleury
Marton Karsai
Hadrien Hours
Sa:i Jouaber
Jacob Levy-Abitbol
Tommaso Venturini