

DATA BREIZH – October 2017  
RENNES, FRANCE

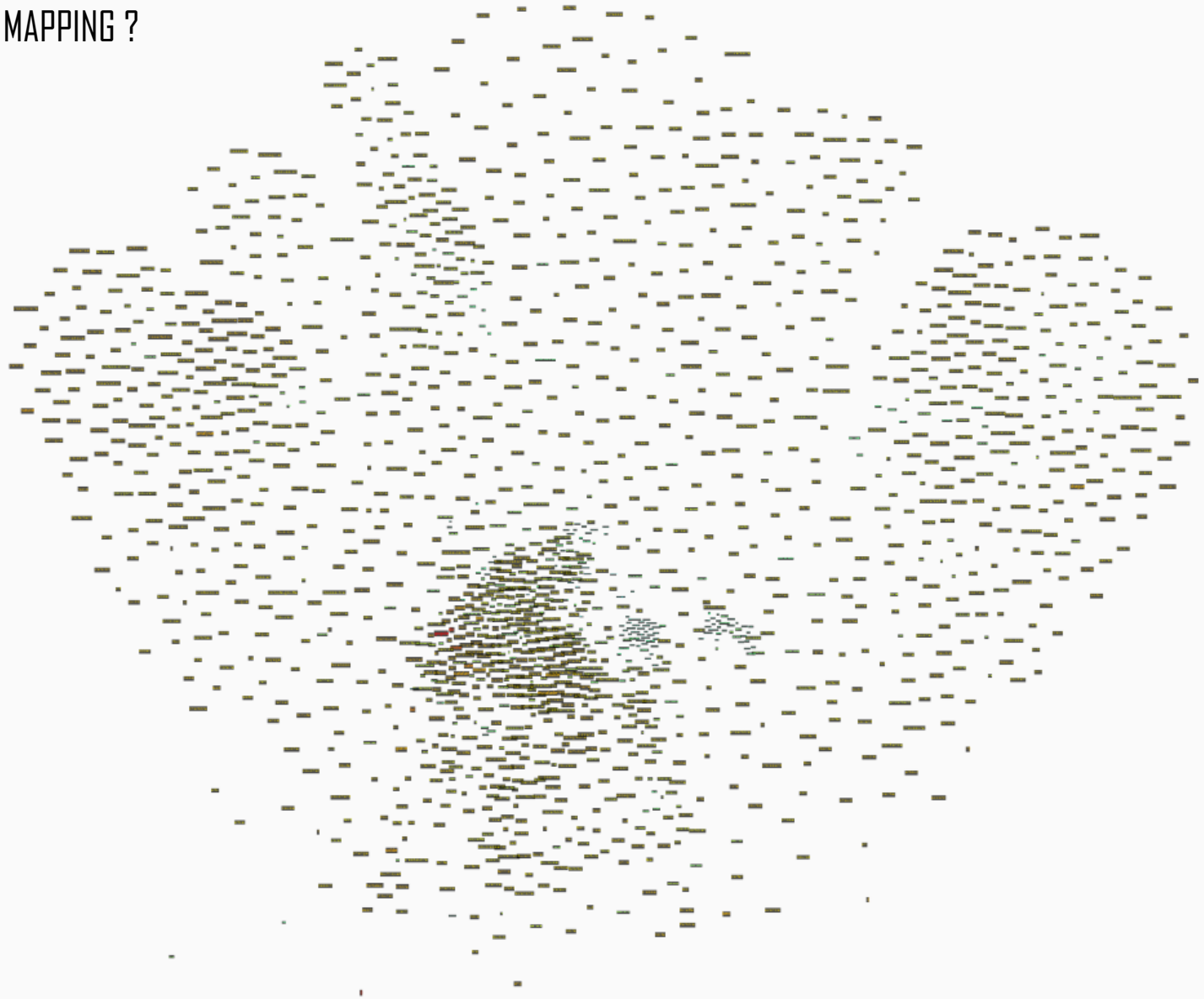
Hidden Social Networks Analysis  
by  
Semantic Mining  
of  
Noisy Corpora

C. Thovex

DATA2B

WE MAKE DATA PRODUCTS

## BRAIN MAPPING ?

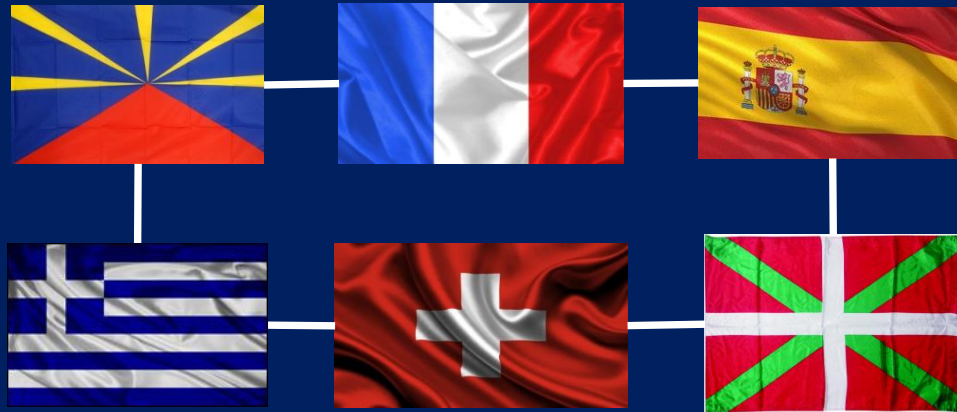


... SOCIAL CONTENT !

519 803 co-occurrences in-between  
1 623 **least common** terms  
shared within 88 522 messages  
by 423 users

Based on a small data sets,  
Intermediate processing may produce big datasets – *e.g.* 88 522 items → 519 803 items.

## 88milSMS\* - French corpus of Short Text Messages, part of the international project SMS4SCIENCE



\*88milSMS. A corpus of authentic text messages in French » Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., Verine B. (2014), produit par l'Université Paul-Valéry Montpellier 3 et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirimm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8 - <http://88milsms.huma-num.fr/index.html>

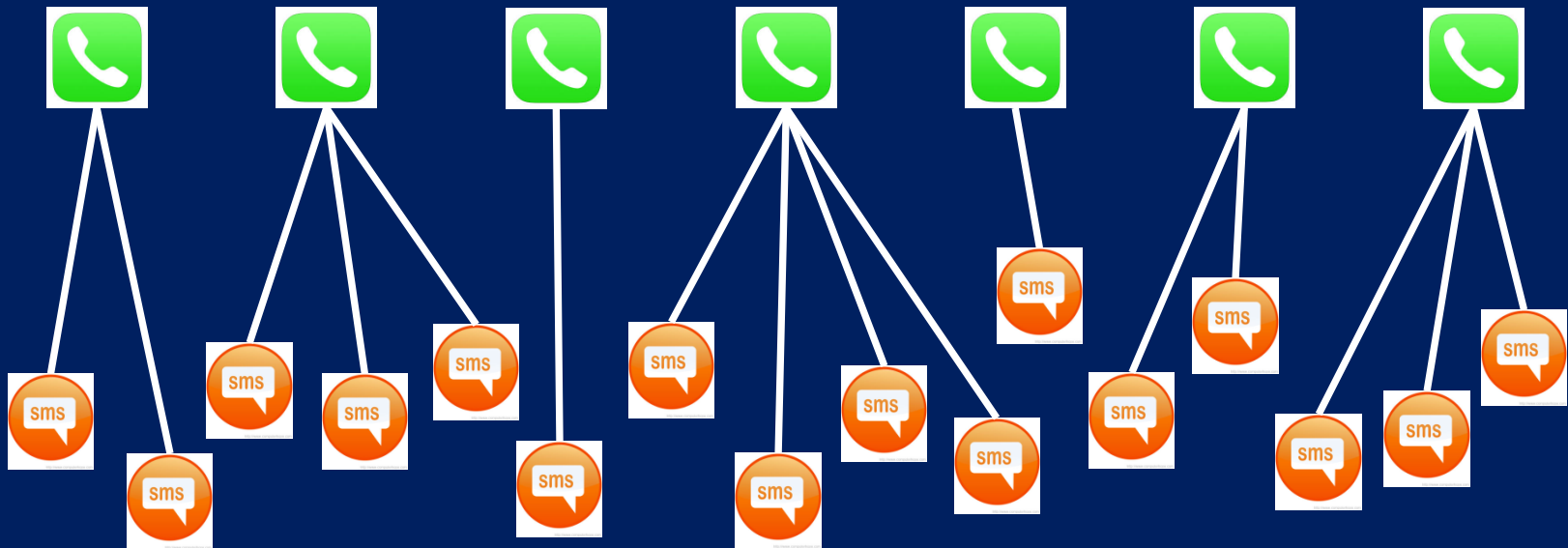
# Small (oriented) data

## Facts table:

Simple facts model : Sender Id (anonymized) , Message Id, Date Time, Text Content

NO destination id → NO relational data (no *user-user* data)

*Who writes to who ? What interests people together ?*



# Lexical similarity, an implicit relation

Refinements of well-known linguistic statistics for retrieving shared vocabulary

TFIDF, Jaccard Index... BM25 [Sparck Jones 1972], [Zaragoza 2004]

- ranking messages for keywords  $(k, k + 1 \dots k_n)$
- Ranking words in message  $m$
- Ranking messages similar to a given message  $m$
- Ranking words similar in-between messages  $(m, m')$

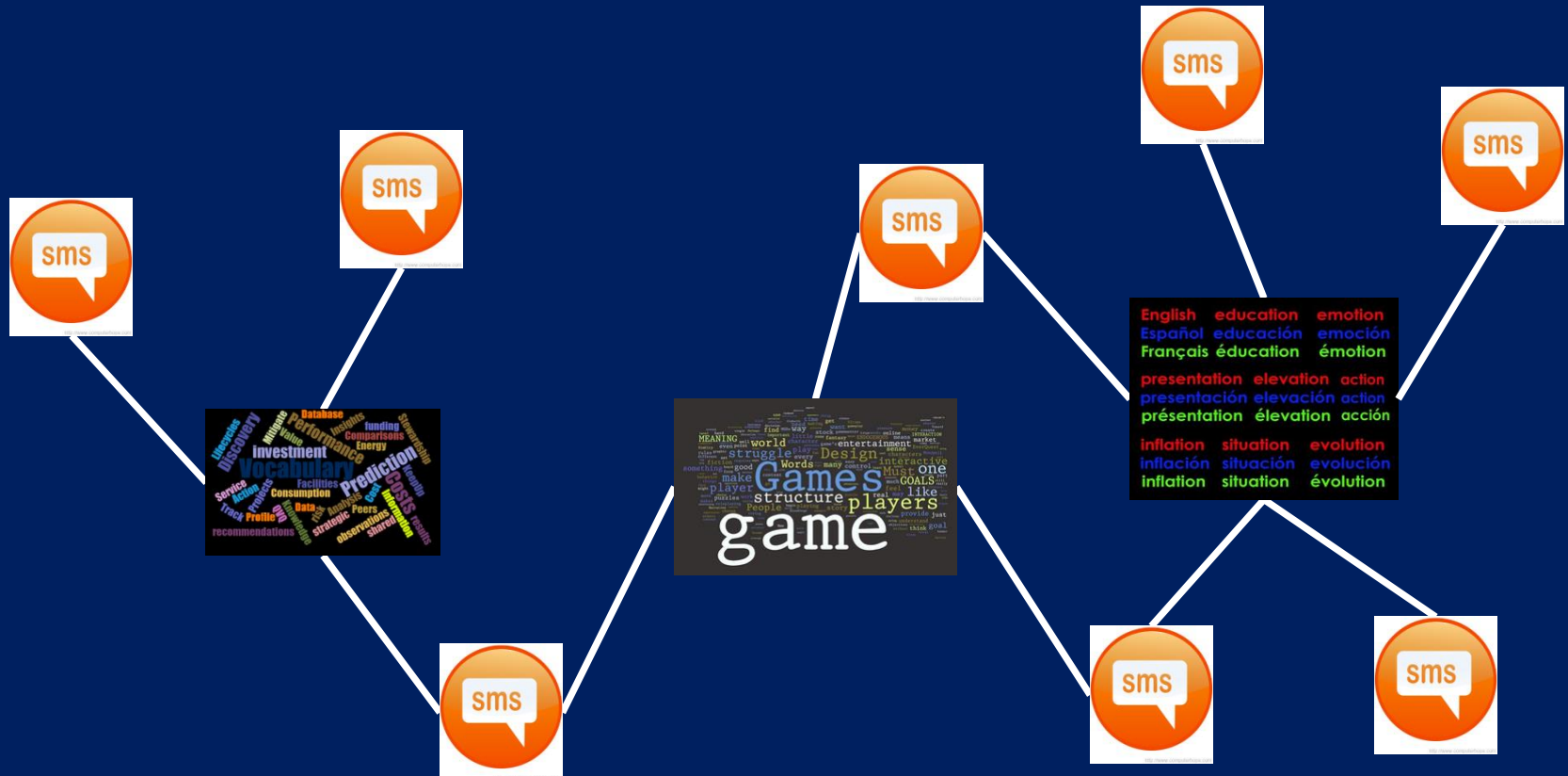
$$Score(w, T) = \frac{2.2 * TF(w, T)}{(0.3 + \frac{0.9 \cdot Dterms}{ADterms} + TF(w, T)) * \frac{8 + QTF(w)}{9 + QTF(w)}} \quad (4)$$

$$Rank(w, T) = Score(w, T) * \log_{10} \left( \frac{Idoc + 0,5}{Tdoc + 0,5} \right) \quad (5)$$



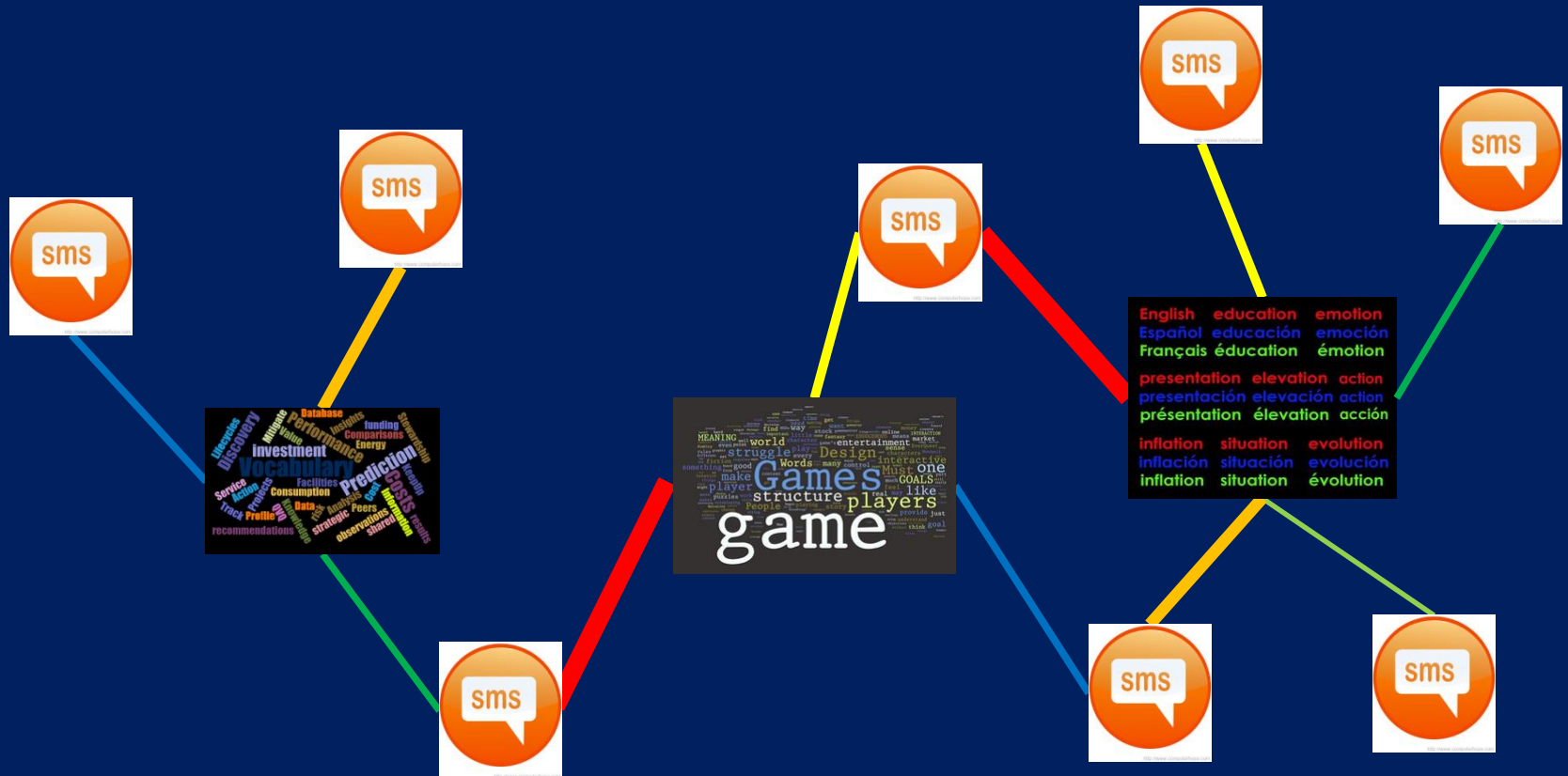
# Lexical similarity for finding semantic ties

Finding social ties within shared vocabulary



# Semantic statistics for ranking social ties

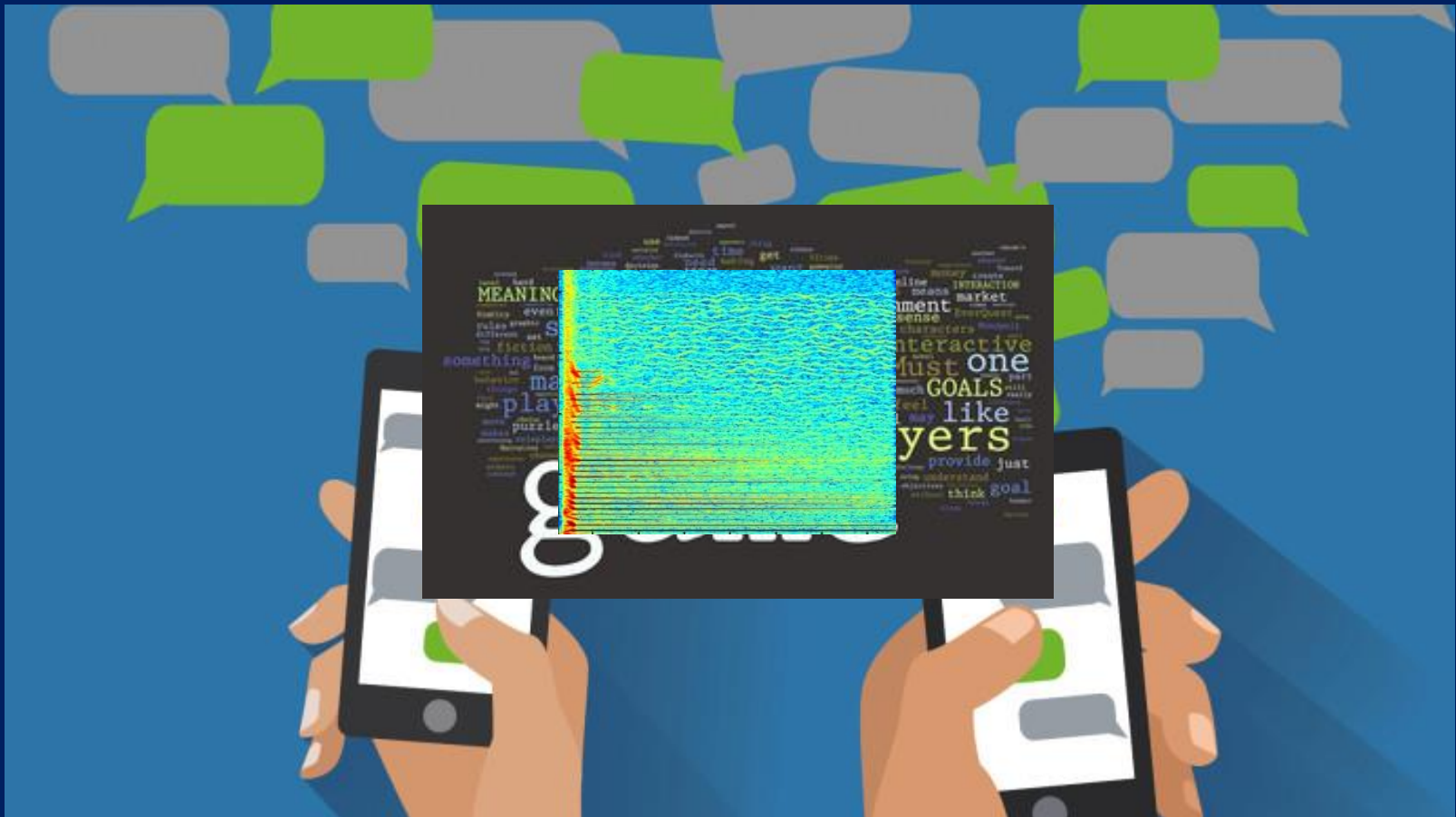
## Ranking social ties found within shared topics





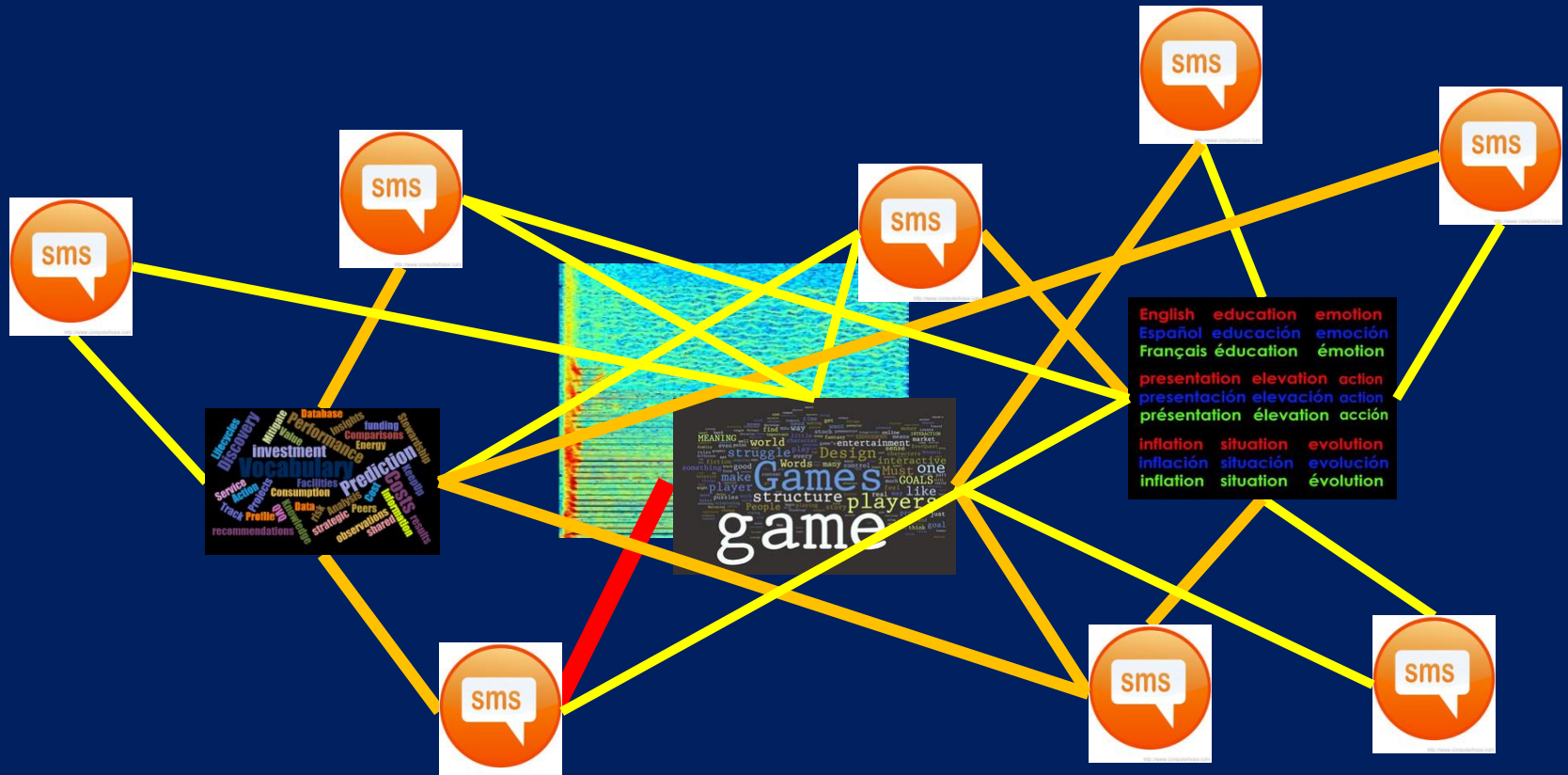
# Semantic statistics in noisy corpora

**Noise reduction** for cleaning and ranking social ties found within shared vocabulary



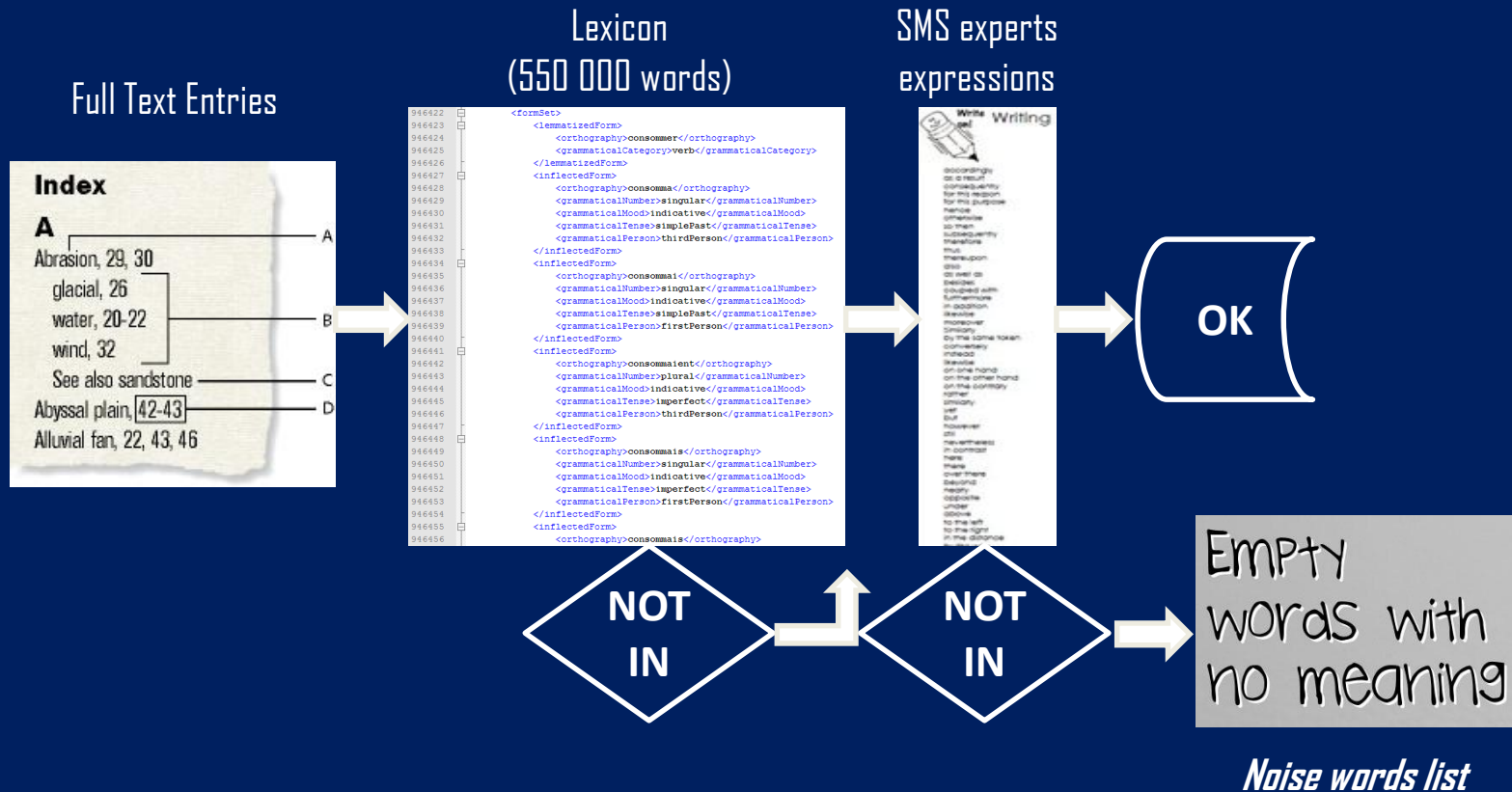
# Linguistic statistics in noisy corpora

How noise affects the extraction of social ties from shared topics



# Dynamic noise reduction

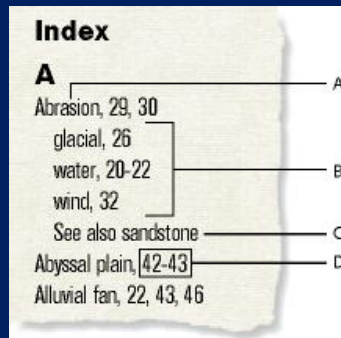
Populating a set of noisy elements **from the studied corpus**



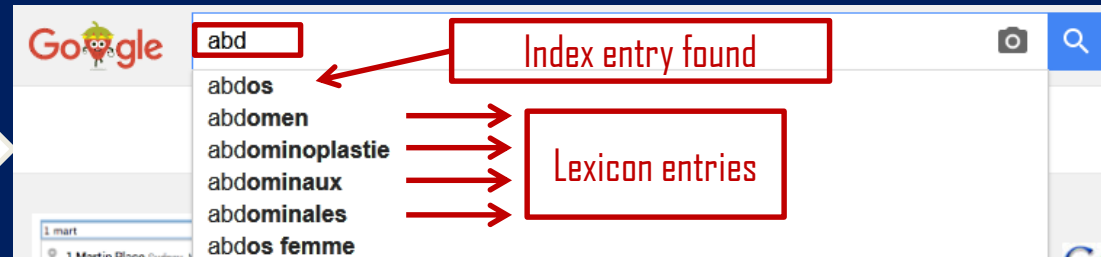
# Semantic noise reduction

Populating a thesaurus from Google knowledge using Google suggest API

Full Text Entries



Algorithm requesting Google suggest API  
example: 'abdos' (familiar abbreviation)



**Thesaurus  
For  
Full Text Search**

# Hidden social networks discovery

Semantic mining and ranking of hidden social graphs

Cleaned  
Full-Text Index

Index	
<b>A</b>	
Abrasion, 29, 30	A
glacial, 26	B
water, 20-22	
wind, 32	
See also sandstone	C
Abyssal plain, 42-43	D
Alluvial fan, 22, 43, 46	

Semantic Full Text Search based algorithm

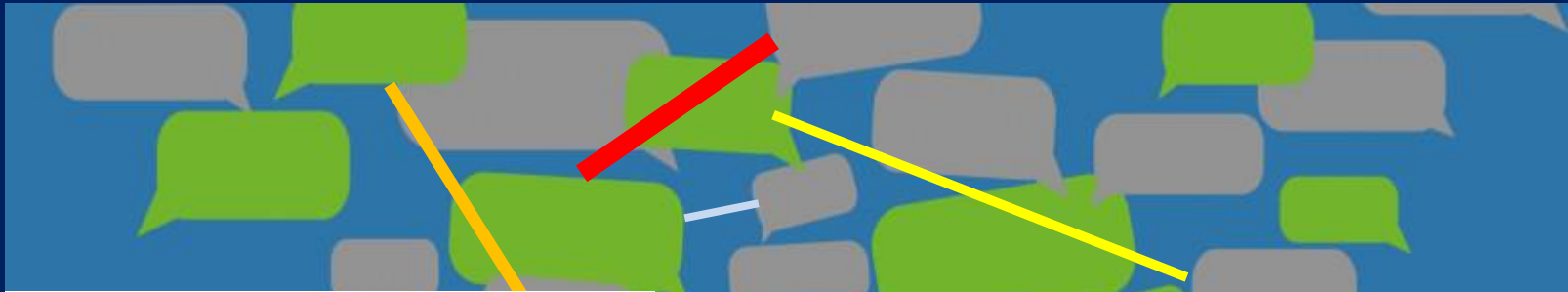


**Hidden Social  
Networks**

-  
**Semantic  
ranking  
of  
nodes and arcs**

# The *knowlinks* metric

A semantic multi-level and distributed metric for social networks analysis



$$J_{(t1,t2)} = |snd : \exists(t1, snd) \wedge (snd, t2)| \quad (13)$$

$$scg_{(t1,t2)} = \frac{\sum_{j=1}^J (sc_{(t1, snd_j)} + sc_{(snd_j, t2)})}{J} \quad (14)$$

$$scvg_{(t1,t2)} = \frac{\sum_{j=1}^J (scv_{(t1, snd_j)} + scv_{(snd_j, t2)})}{J} \quad (15)$$

$$agg_{(t1,t2)} = \frac{\sum_{j=1}^J (avgrk_{(t1, snd_j)} + avgrk_{(snd_j, t2)})}{J} \quad (16)$$

$$know_{(t1,t2)} = 10 \cdot scg_{(t1,t2)}^2 + scvg_{(t1,t2)} + \log(agg_{(t1,t2)}) \quad (17)$$

$$knowlink_{(t1,t2)} = \frac{\sum_{j=1}^J know_{(t1,t2)}}{J} \quad (18)$$

$J$ : count of concurrent terms in sender texts

*scg*: average inverse of sum of ranks of a couple of concurrent indexed terms in sender messages

*scvg*: average of sum of ranks of a couple of concurrent indexed terms in sender messages

*agg*: average of average ranks of a couple of concurrent indexed terms in sender messages

*know*: complex ranking of a couple of concurrent indexed terms in sender messages (separation improvement)

*Knowlink*: Average *know* metric of a couple of concurrent indexed terms in sender messages



# Semantic ties in hidden social networks

Data model and facts table

Full Text Search from clean index keywords:  
relation 'keyword - ID\_NumTel'

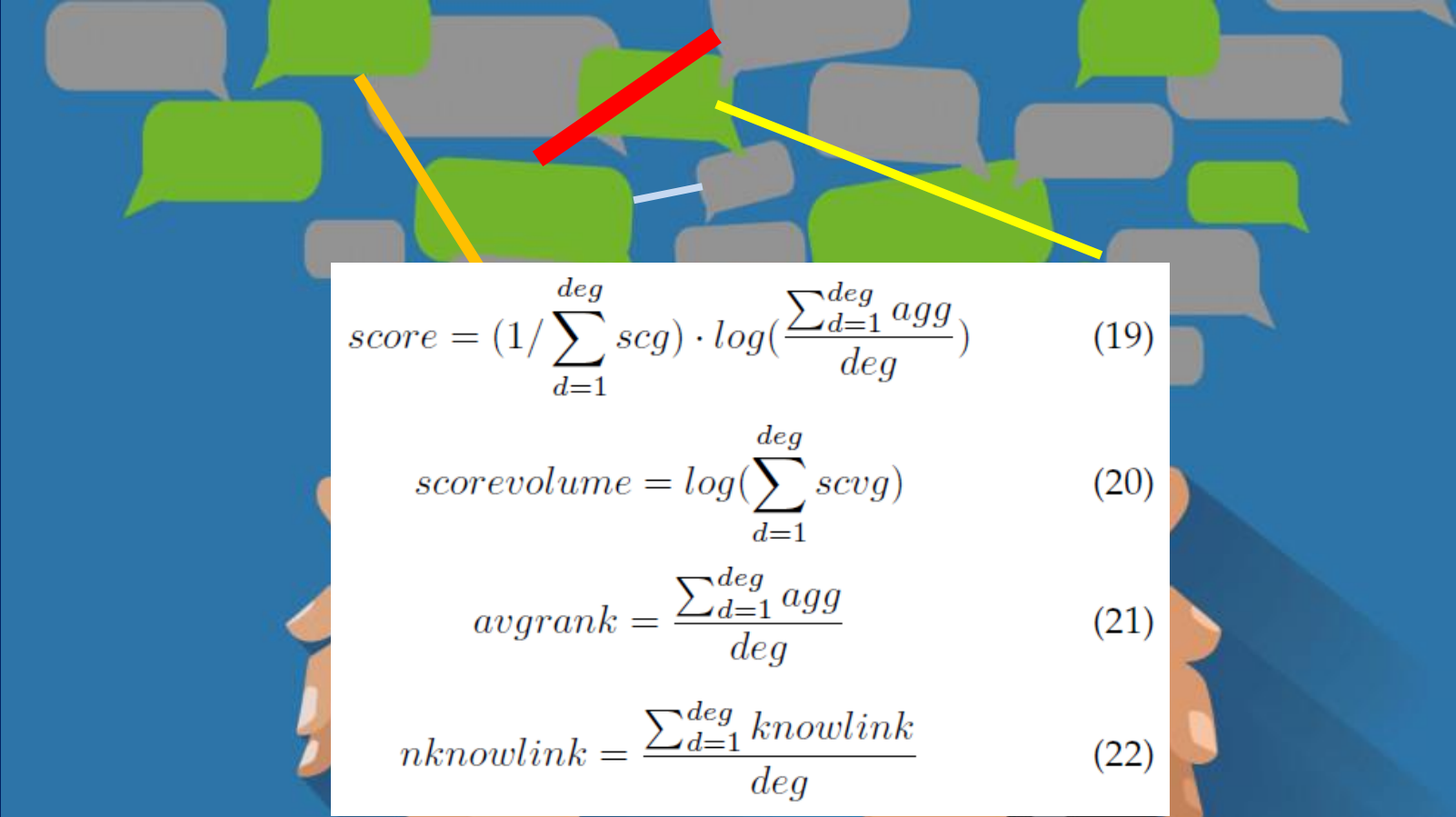
:(	101	0,161975161298183	0,161975161298183	6,17378610390194	480
:(	105	0,161975161298183	0,161975161298183	6,17378610390194	480
:(	11	0,13476037553766	0,13476037553766	7,4205789054108	417
:(	110	0,161224389159705	0,161224389159705	6,20253551718792	247
:(	120	0,145319027721925	0,145319027721925	6,88141130364254	324
:(	124	0,142301870508587	0,142301870508587	7,02731451403978	225
:(	134	0,145319027721925	0,145319027721925	6,88141130364254	324

Semantic mining of social ties

124614	arabe	revient	0,286734804986887	0,286734804986887	13,9509431421782	428	92,7521680321015
124615	patte	revenant	0,286734804986887	0,286734804986887	13,9509431421782	428	92,7521680321015
124616	arabe	revenu	0,286734804986887	0,286734804986887	13,9509431421782	428	92,7521680321015
124617	arabe	revenant	0,286734804986887	0,286734804986887	13,9509431421782	428	92,7521680321015
124618	angleterre	disant	0,276142370630164	0,276142370630164	14,8985593614633	303	92,7518482641462
124619	angleterre	dire	0,276142370630164	0,276142370630164	14,8985593614633	303	92,7518482641462
124620	angleterre	dit	0,276142370630164	0,276142370630164	14,8985593614633	303	92,7518482641462
124621	pardon	relation	0,324053044786714	0,324053044786714	13,479620157128	446,6666...	92,7511537591818
124622	boucle	chouchoute	0,297792813311868	0,297792813311868	13,4804433338506	504	92,7511427383191
124623	mordant	projet	0,282574036399993	0,282574036399993	14,3875133097106	362	92,750918118461
124624	mordu	projet	0,282574036399993	0,282574036399993	14,3875133097106	362	92,750918118461
124625	guitare	mat	0,280072286343286	0,280072286343286	14,2936385946534	380	92,7507096344911

# Semantic flows coherence in nodes weighting

Coherent weighting degree of semantic flows in hidden social networks


$$score = (1 / \sum_{d=1}^{deg} scg) \cdot \log\left(\frac{\sum_{d=1}^{deg} agg}{deg}\right) \quad (19)$$

$$scorevolume = \log\left(\sum_{d=1}^{deg} scvg\right) \quad (20)$$

$$avgrank = \frac{\sum_{d=1}^{deg} agg}{deg} \quad (21)$$

$$nknowlink = \frac{\sum_{d=1}^{deg} knowlink}{deg} \quad (22)$$

# Semantic weights in hidden social networks

## A linear metric and long tail distribution

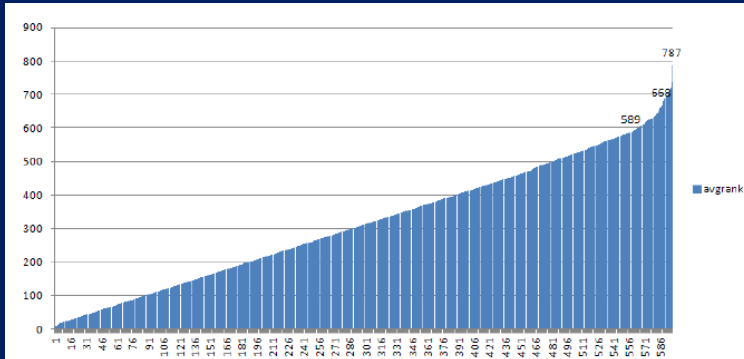


Fig. 1. Arcs weights *avgrk* - ordered values

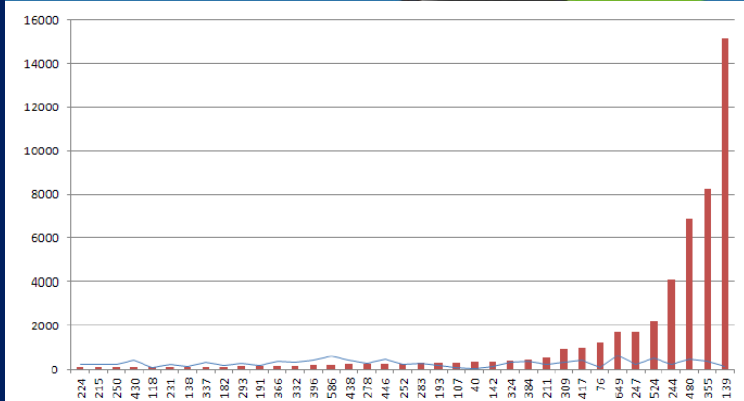


Fig. 2. Arcs distribution in *avgrk* - count > 99 (head)

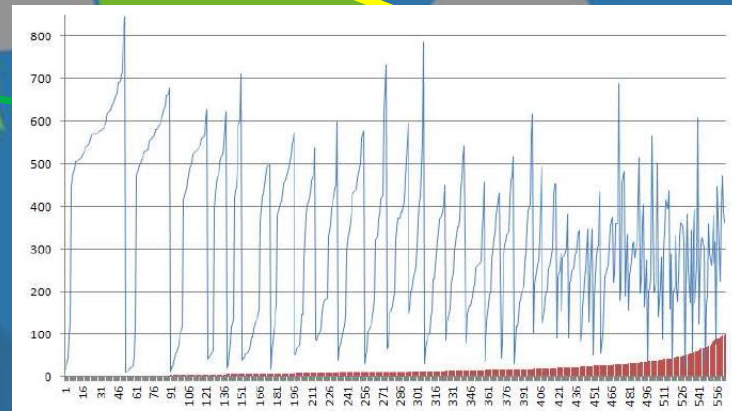


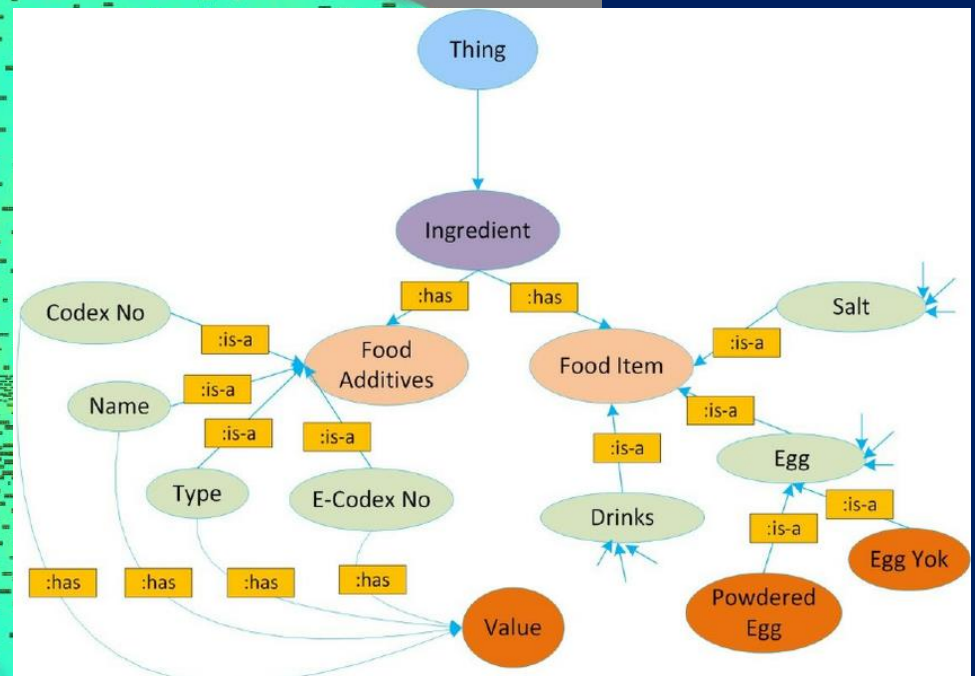
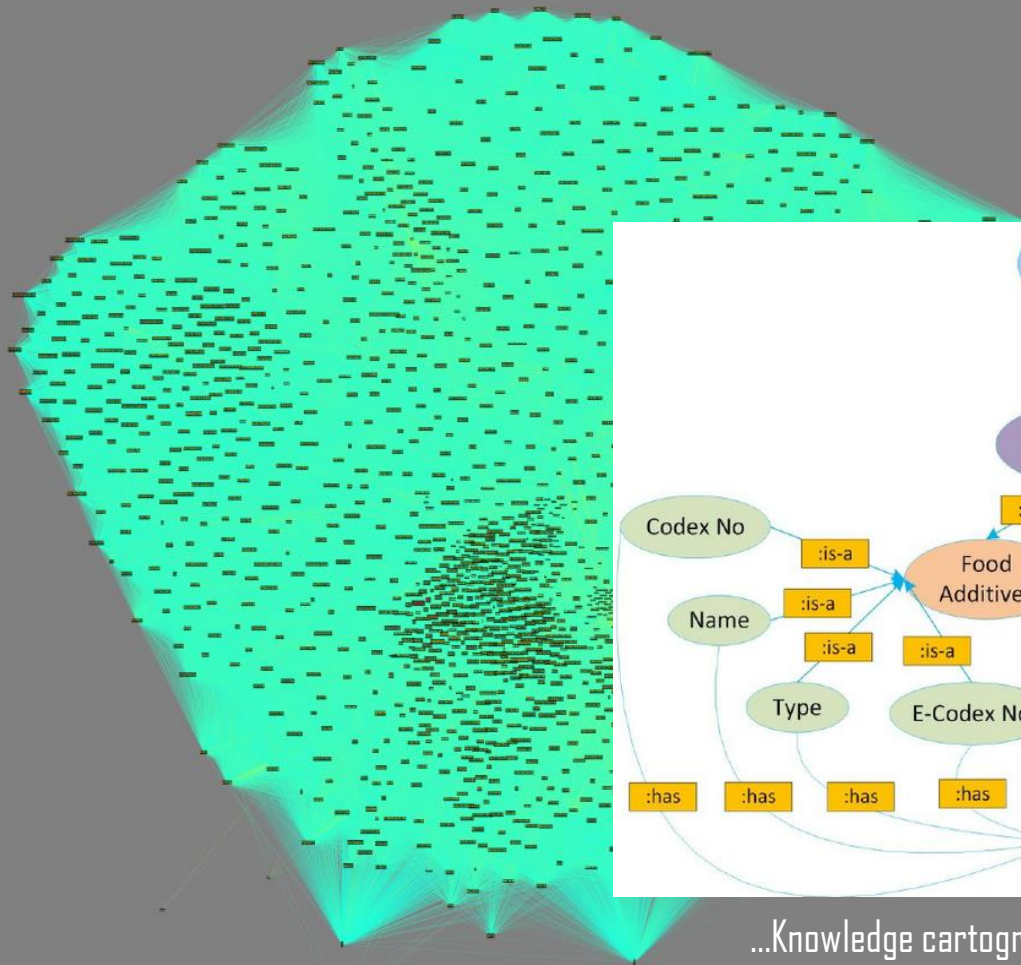
Fig. 3. Arcs distribution in *avgrk* - count < 100 (tail)

# Hidden social networks **analysis**

Social networks analysis with semantic weight within directed social graphs



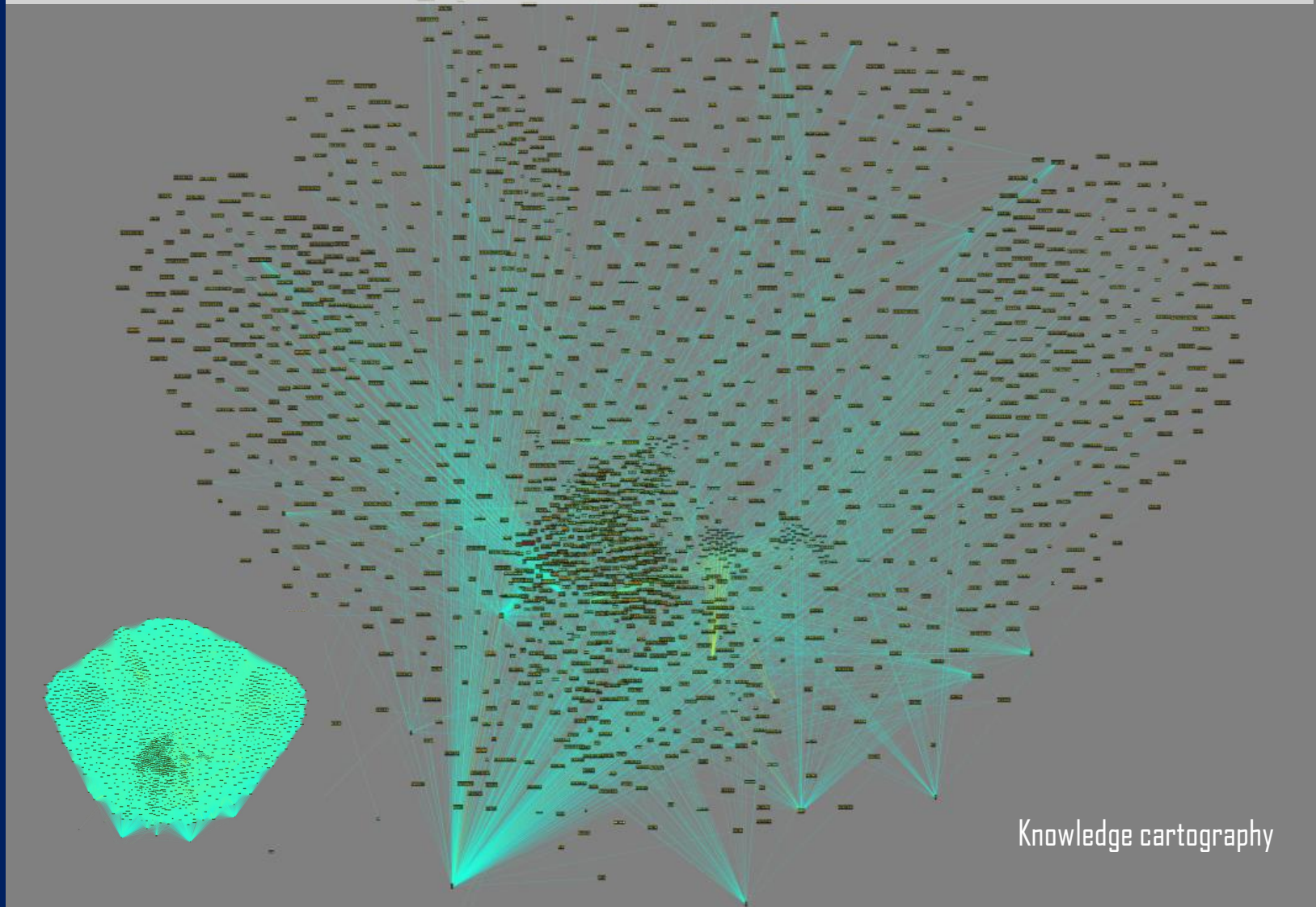
# Knowledge cartography, automatic ontology building, mind mining ? ...



...Knowledge cartography : towards ontology building



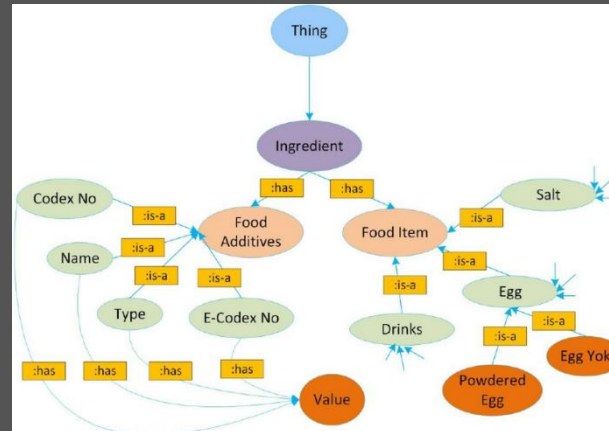
# Spanning tree of least common terms...



Knowledge cartography

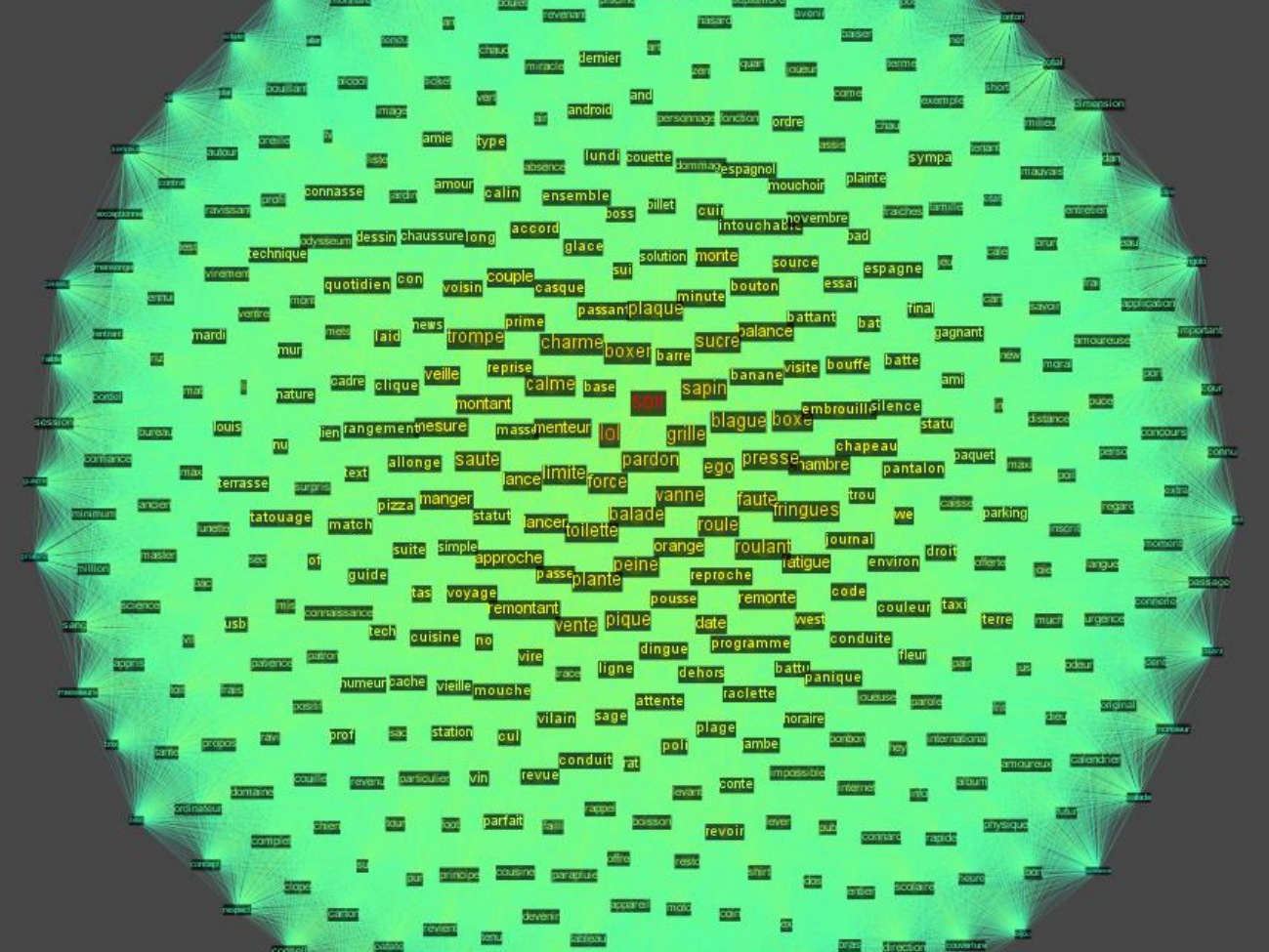


# Spatialized spanning tree (Yifan Hu's algorithm)...



Organized knowledge cartography

# The most common terms as a *knowledge sphere*

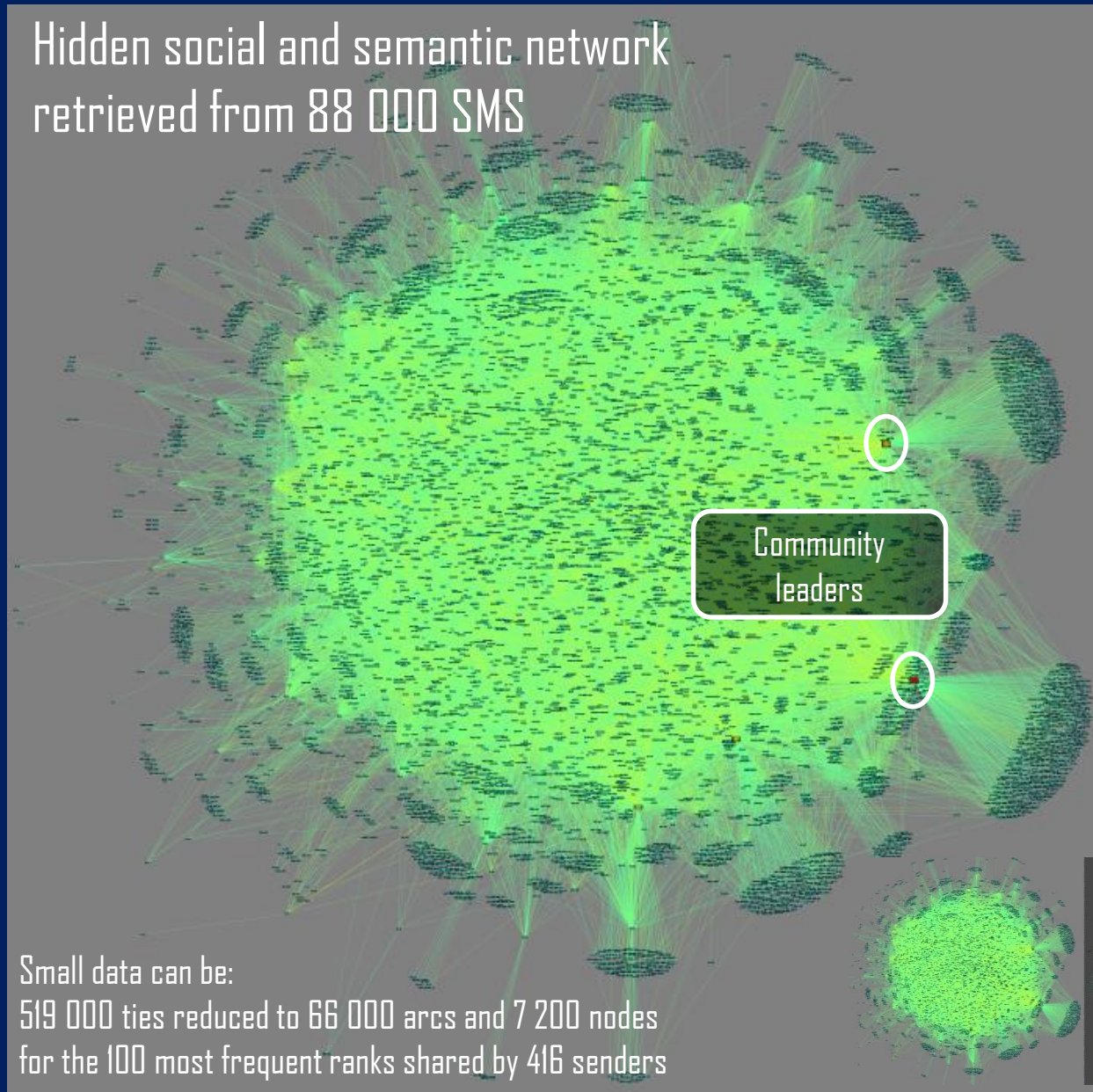


Talk of the studied crowd  
, wisdom of crowd ? [Surowiecki2005]

# The most common terms as a *knowledge sphere*

Talk of the studied crowd  
, wisdom of crowd ? [Surowiecki2005]

# Hidden social and semantic network retrieved from 88 000 SMS

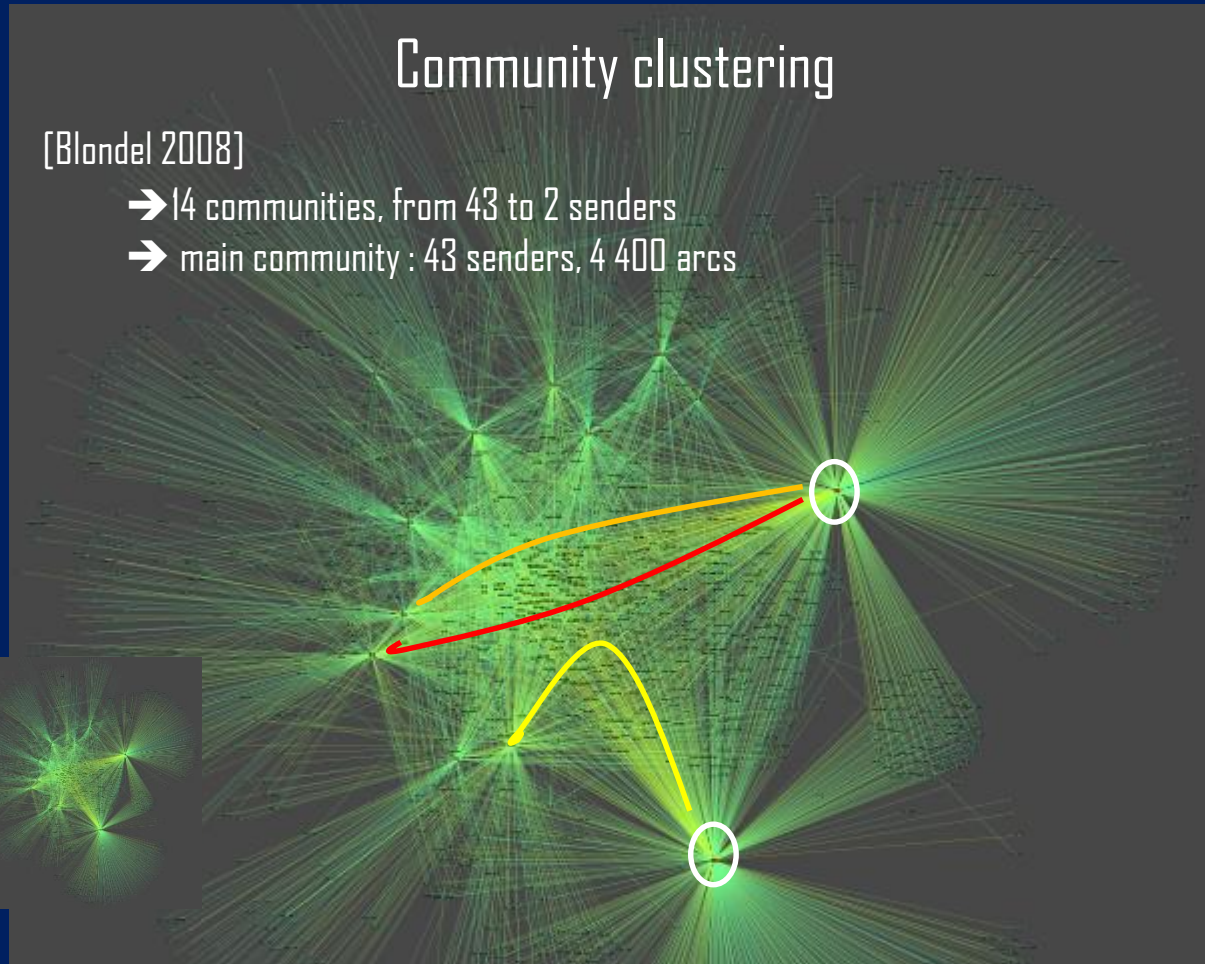
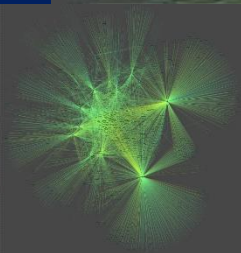
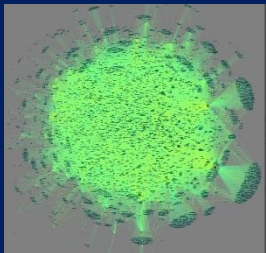




# Community clustering

[Blondel 2008]

- 14 communities, from 43 to 2 senders
- main community : 43 senders, 4 400 arcs



Semantic ranking *scv*

Probabilistic ties associating users to senders (relationships retrieval)

- OK for finding pairs (sender, user) thanks to the semantic convergence expressed by knowlinks, and/or edge betweenness and/or other classification methods such as random tree forest (gradient boosting).

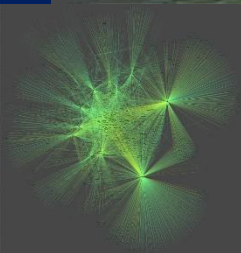
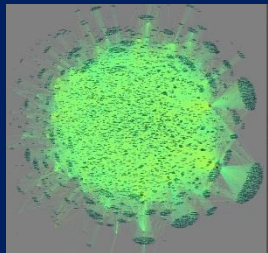
# Community clustering

[Blondel 2008]

- 14 communities, from 43 to 2 senders
- main community : 43 senders, 4 400 arcs

Shared  
vocabulary  
proper to the  
community

Shared  
vocabulary  
proper to  
community  
leaders



Semantic ranking *scv*

enables to show the **influence of community leaders**  
on the **community shared vocabulary** – yellow arcs

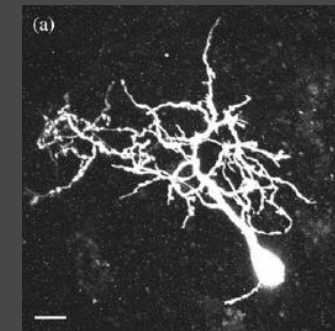


Filter: top 1 *knowlink* arcs in frequent terms sub-graph

Not an ontology designed by a geek ...

Nor dendritic cells...

A representation of parts of speech in the talk of crowd



<https://www.researchgate.net>



[illegible]

## Perspectives and questions

- Scalability: experimenting with larger datasets – Big Data
- Extension: *knowlinks* as a premise for *deep social learning* – sentiment analysis [1]
- Transdisciplinarity (what if ) :
  - use in knowledge engineering/ontology building ?
  - use in social psychology ?
  - use in neuroscience, neuro-marketing ?

[1] *Hidden Social Networks Analysis by Semantic Mining of Noisy Corpora*. In The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASDNAM 2016, San Francisco, CA, USA., pages 868–875, 2016.

DATA BREIZH – October 2017  
RENNES, FRANCE

Christophe Thovex

Dir. of R & D - DATA2B

Associated Scientist - CNRS

[christophe.thovex@data2b.net](mailto:christophe.thovex@data2b.net)

*FPGAs mimic brain plasticity for computers*



DATA2B

WE MAKE DATA PRODUCTS