

Where the **dead blogs** are

A Disaggregated Exploration of Web archives to Reveal Extinct Online Collectives



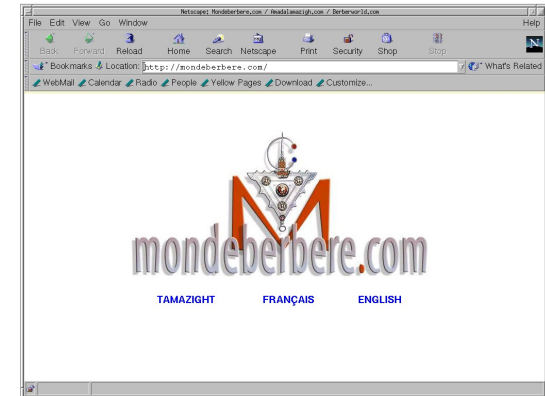
Quentin Lobbé (LTCI, Télécom ParisTech, Université Paris Saclay & Inria)

The **online** representations of diasporas

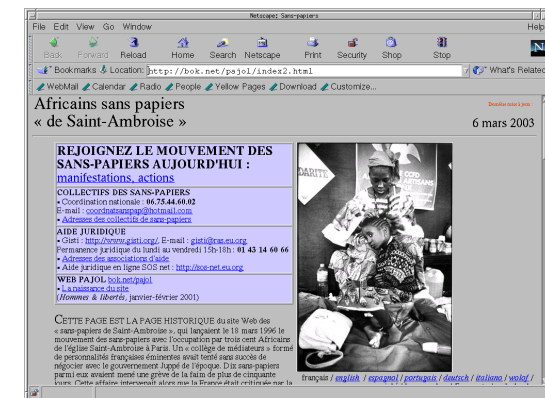
> Migrants are the actors of a culture of bonds



> Personal laptop of a couple of Philippines workers in Paris, Diminescu, D. (2005)



> mondeberbere.com, Morocco, 2002



> bok.net/pajol, France, 1996

> By the mid 2000's, sociologists started to study the many digital traces left by diasporas

The e-Diasporas Atlas (1/2)

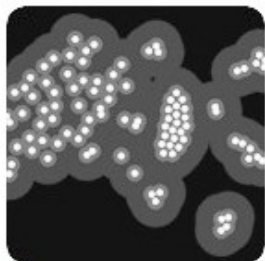
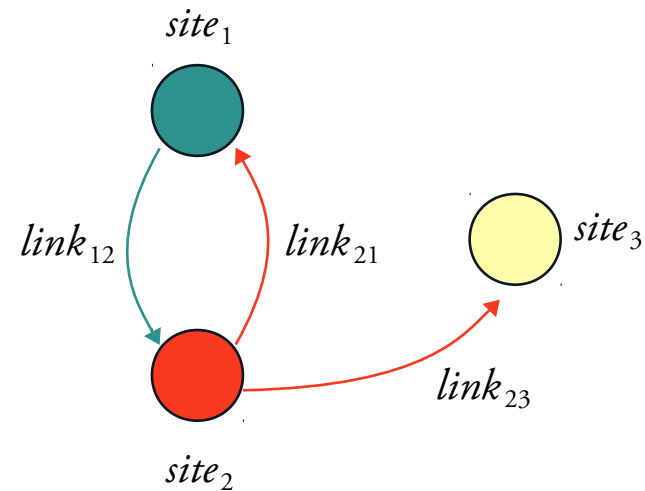
> A multidisciplinary effort to discover and study online migrant collectives

A **migrant web site** is a Web site created or managed by migrants and/or that deals with them

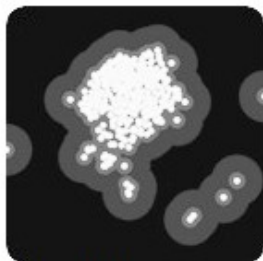
An **e-Diaspora** is a directed network of migrant Web sites linked by url (hypertext links)

An e-Diaspora is both **online** and **offline**

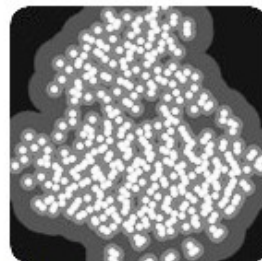
10.000 migrant Web sites crawled, categorized and organized among **30** e-diasporas



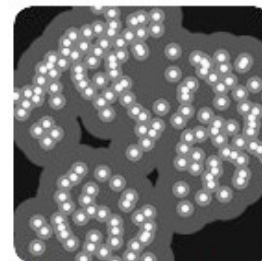
Lebanese corpus



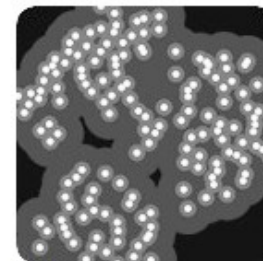
Macedonian corpus



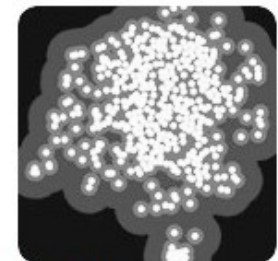
Mexican corpus



Moroccans on FB



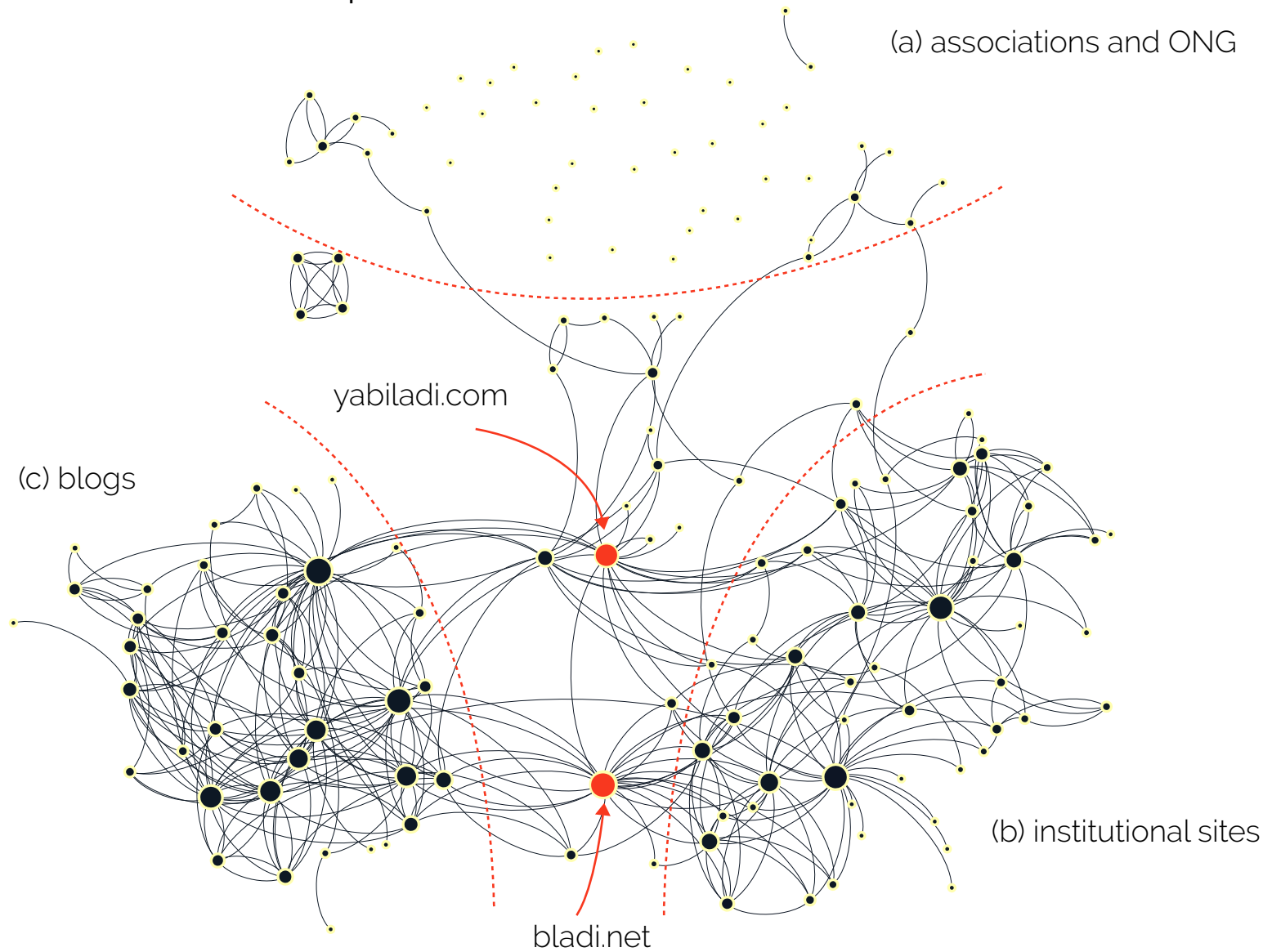
Moroccan corpus



Nepali corpus

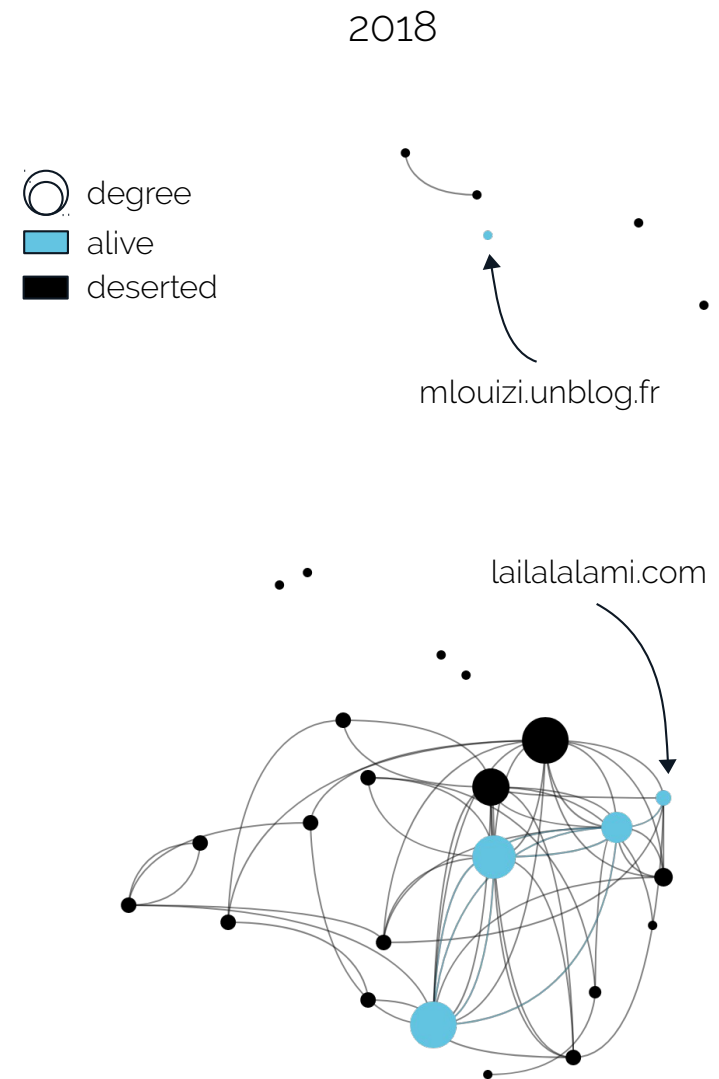
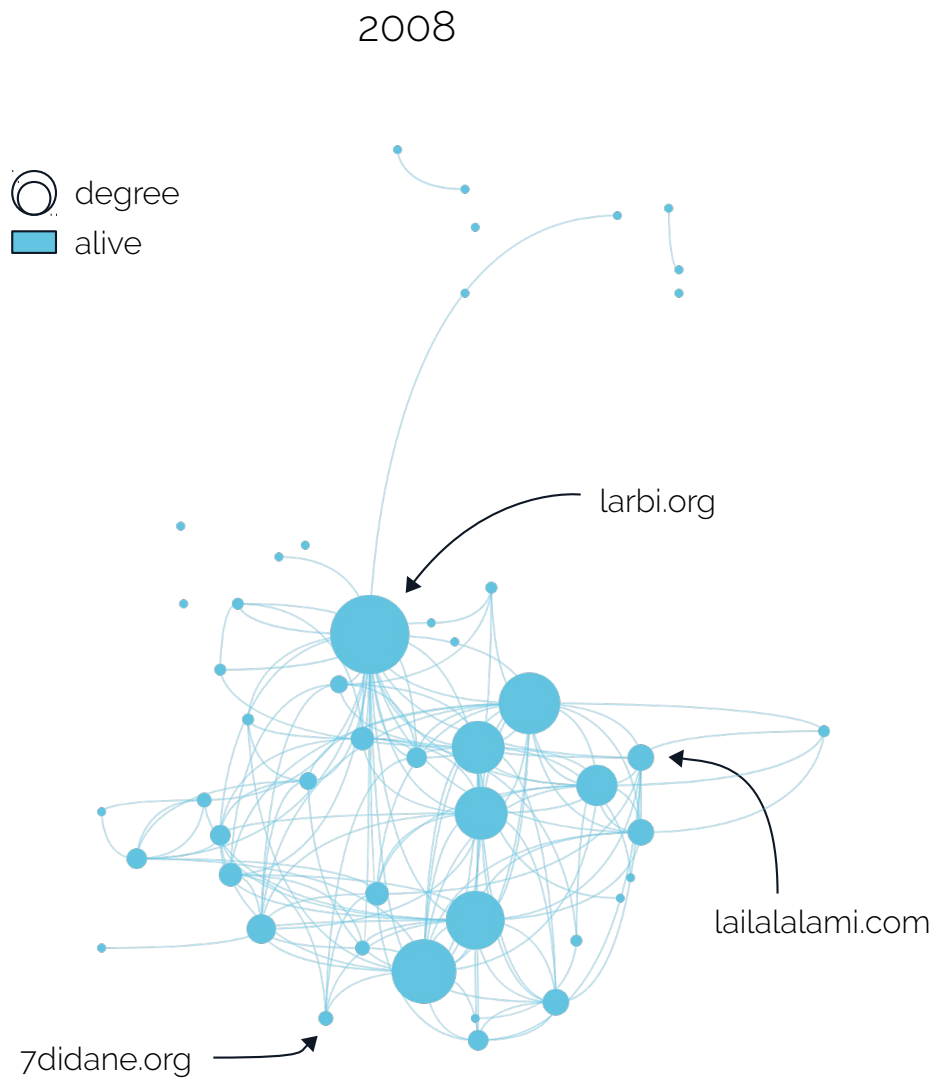
The e-Diasporas Atlas (2/2)

> How to read and use the map?



The question of **extinct** online collectives

> A community for which too few or incomplete traces remain on the living Web

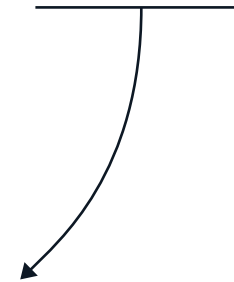


> The Moroccan blogosphere (close up and evolution)

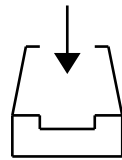
> What happened to the dead Moroccan blogs?

We hypothesize that the structure of the blogosphere is **permeable** to the impact of exogenous **events** or **shocks** such as political or social mobilisations.

We will conduct an exploration of the e-Diasporas corpus of **Web archives** to find their remaining archived traces.



The e-Diaspora Atlas is also a corpus of Web archives

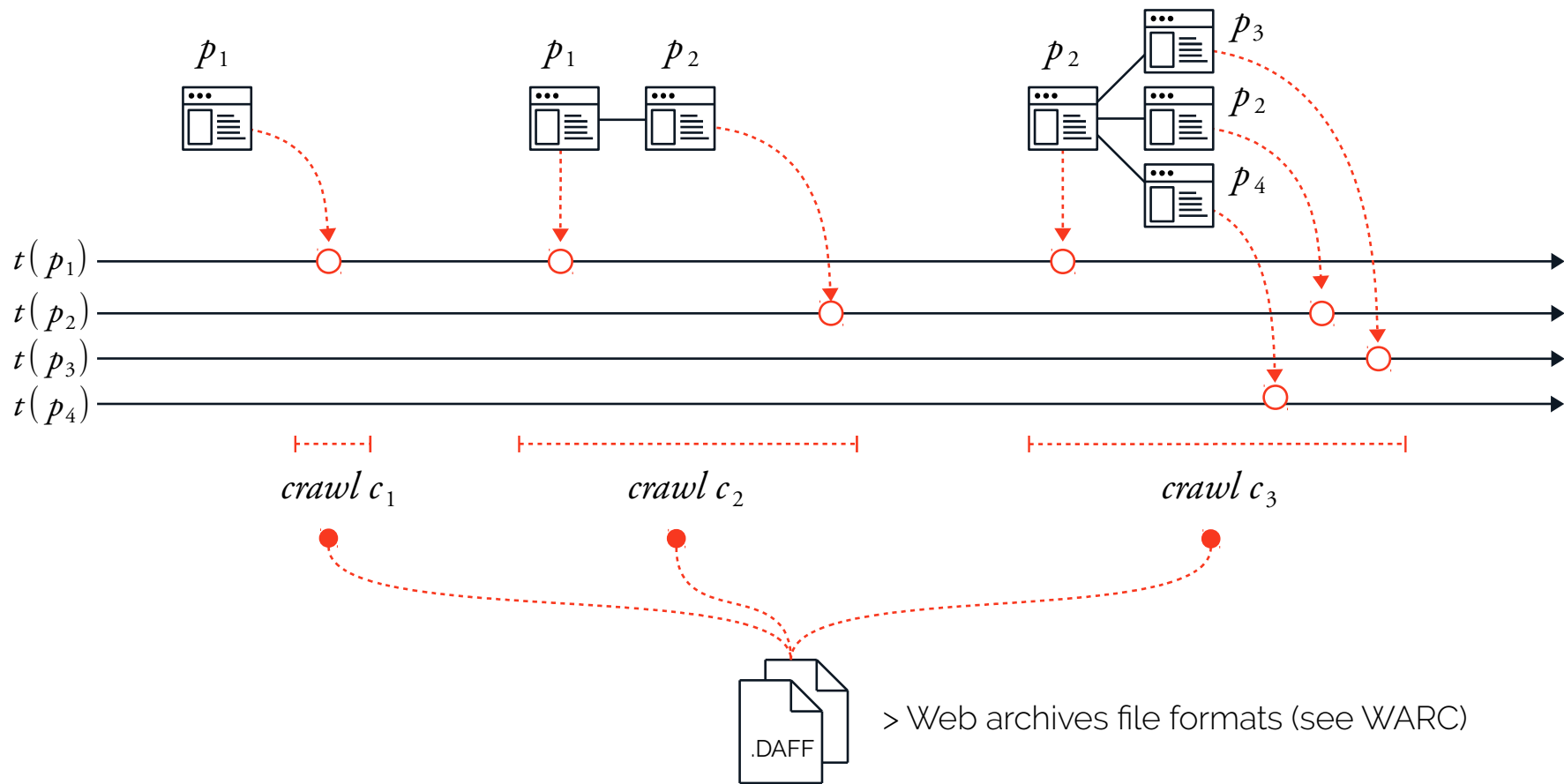


1030 M of Web pages
70 TB
Crawled weekly or monthly (2010-2014)
Hosted and performed by the INA

Archiving the Web? (1/2)

> The preservation of our digital heritage

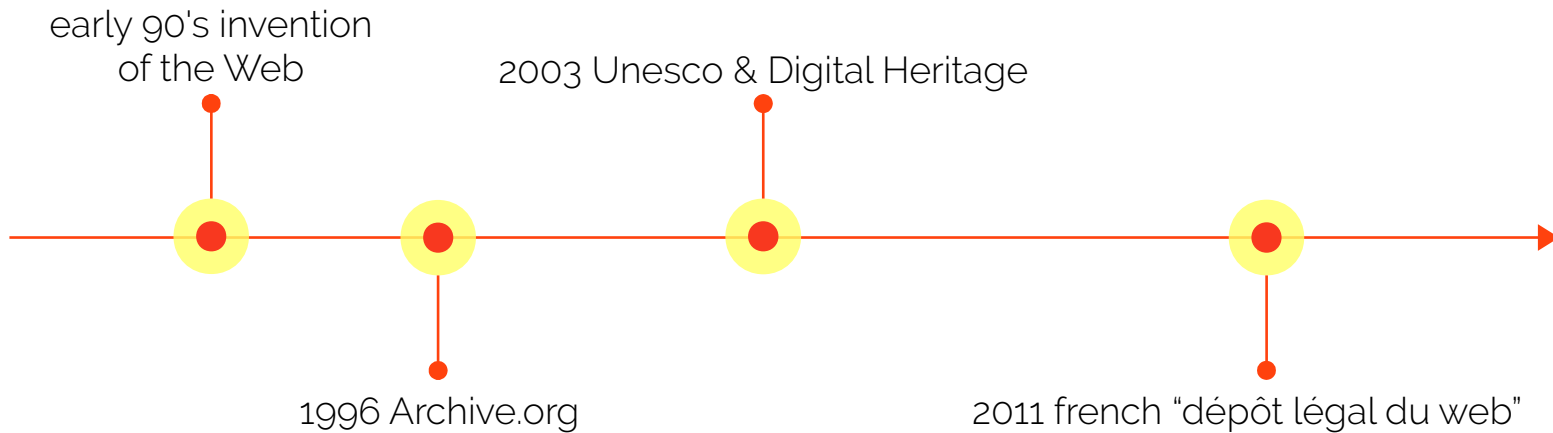
From the continuous Web



To a discrete corpus of Web archives

Archiving the Web? (2/2)

> Exploration tools are designed for manual and focused analysis



> search by URL

WayBackMachine

> full text


ARQUIVO.PT

> aggregators



> local access



> Why is it so hard to conduct an exploration of Web archives at scale ?

Web archives are **not direct traces** of the Web (1/2)

> Web archives are direct traces of the crawler

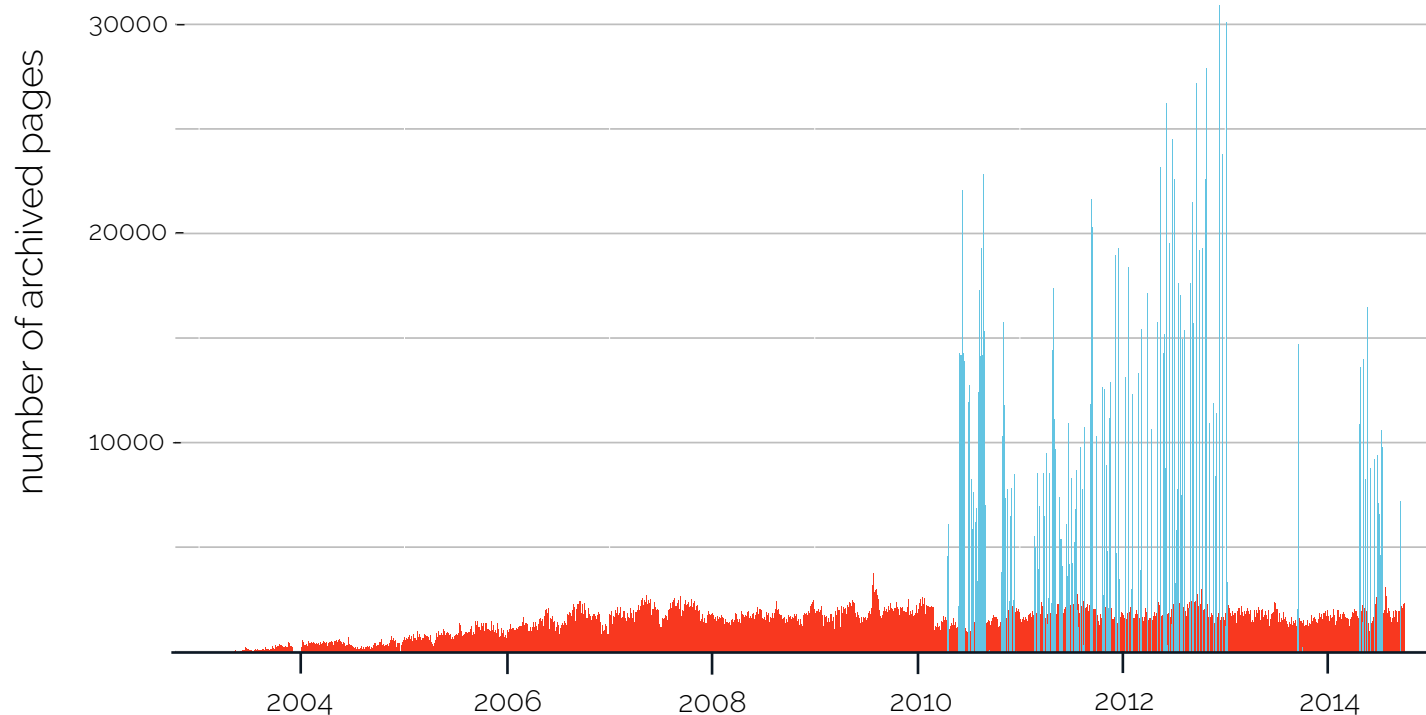
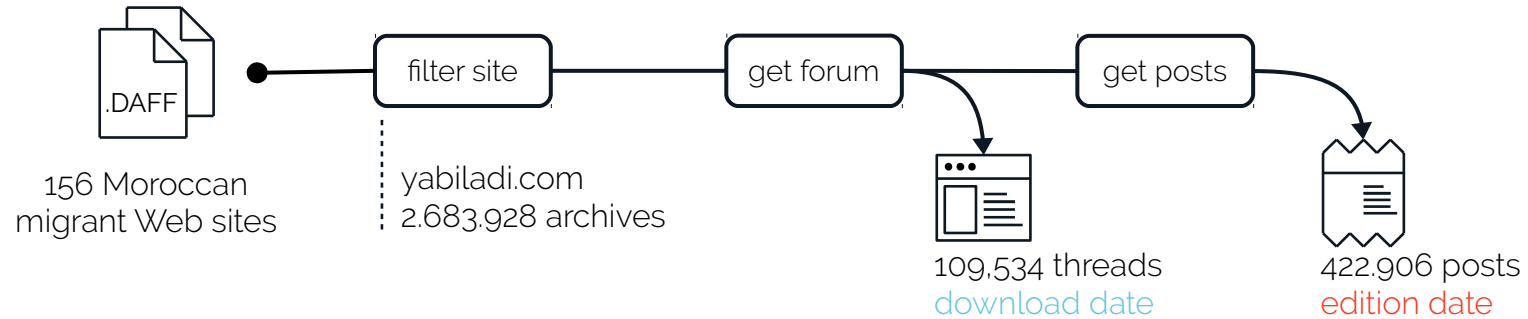


> "Boulevard du Temple", Louis Daguerre, 1838

> Web archives are built on top of Web pages and induce **crawl legacy effects**

Web archives are **not direct traces** of the Web (2/2)

> Going under the level of a Web page



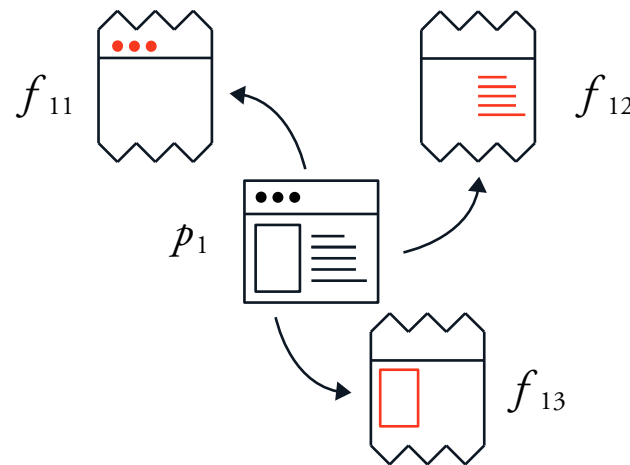
In order to conduct a large scale exploration of the Web that was:

- > We propose to introduce a **new unit of exploration** of Web archives corpora to avoid all kind of crawl legacy effects and maximise the historical accuracy of our forthcoming exploration.

The Web **fragment** (1/3)

> Definition

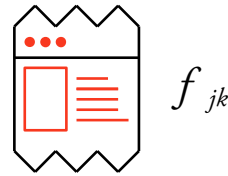
Considering the Web page as the unit of access and consultation to the Web, built using it's own writing modalities and noticing that from the point of view of human perception, a Web page is the result of a logical arrangement of distinct semantic components. We define **the Web fragment as a semantic and syntactic subset of a given Web page.**



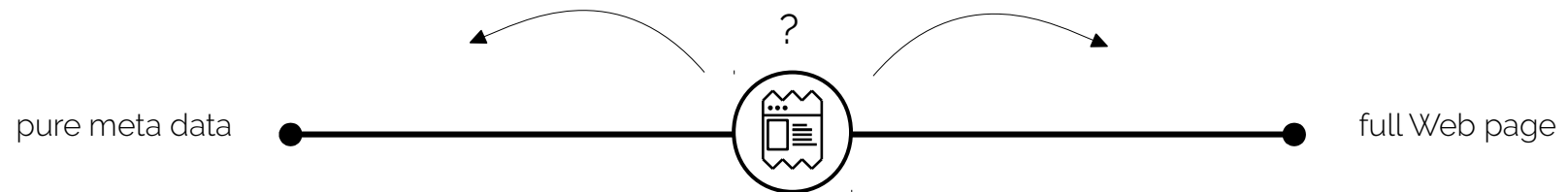
The Web fragment (2/3)

> Definition

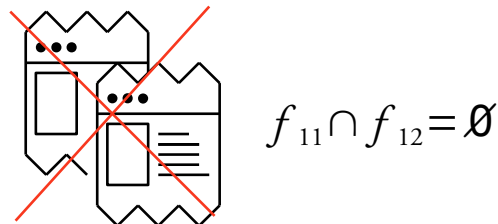
It's a coherent and self sufficient set of textual, visual or audio content



There is a scale relationship between a Web page and its fragments



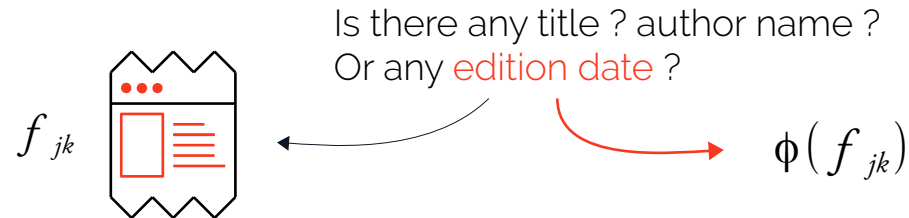
Within the same Web page, two Web fragments cannot overlap



The Web **fragment** (3/3)

> Definition

It goes with an associated set of categorised informations



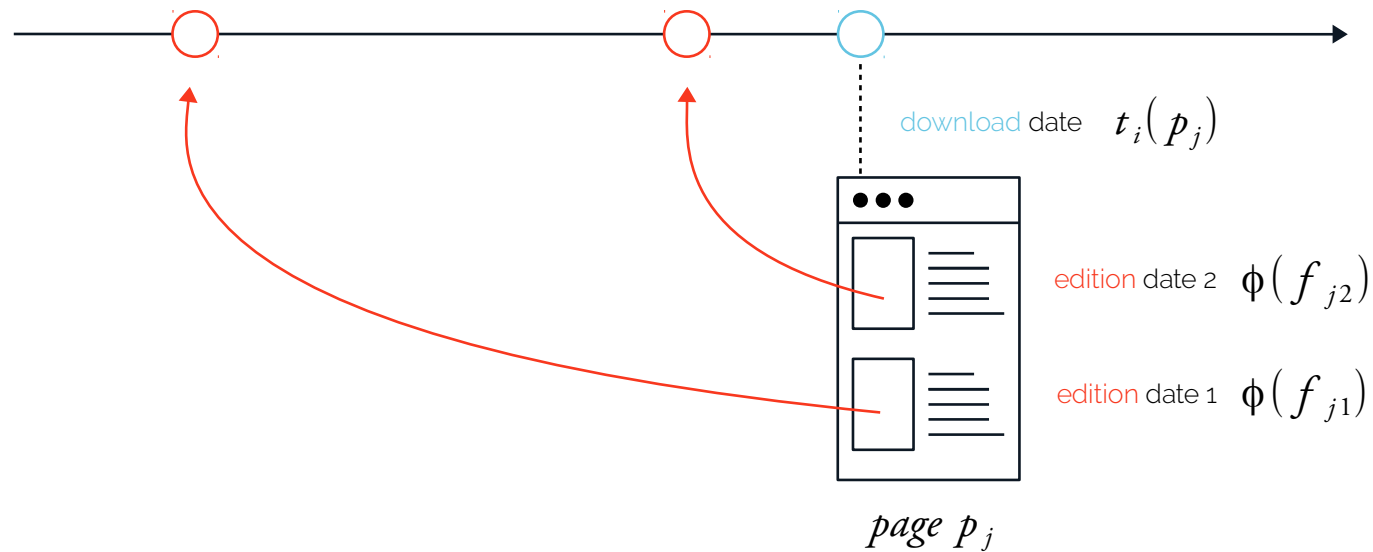
It encompass the writing and sharing elements used for publishing and sharing its content



Upscaling the exploration (1/3)

> Crawl blindness

$$\forall p_j, f_{jk} \exists \phi(f_{jk}) : \phi(f_{jk}) \leq t_i(p_j)$$



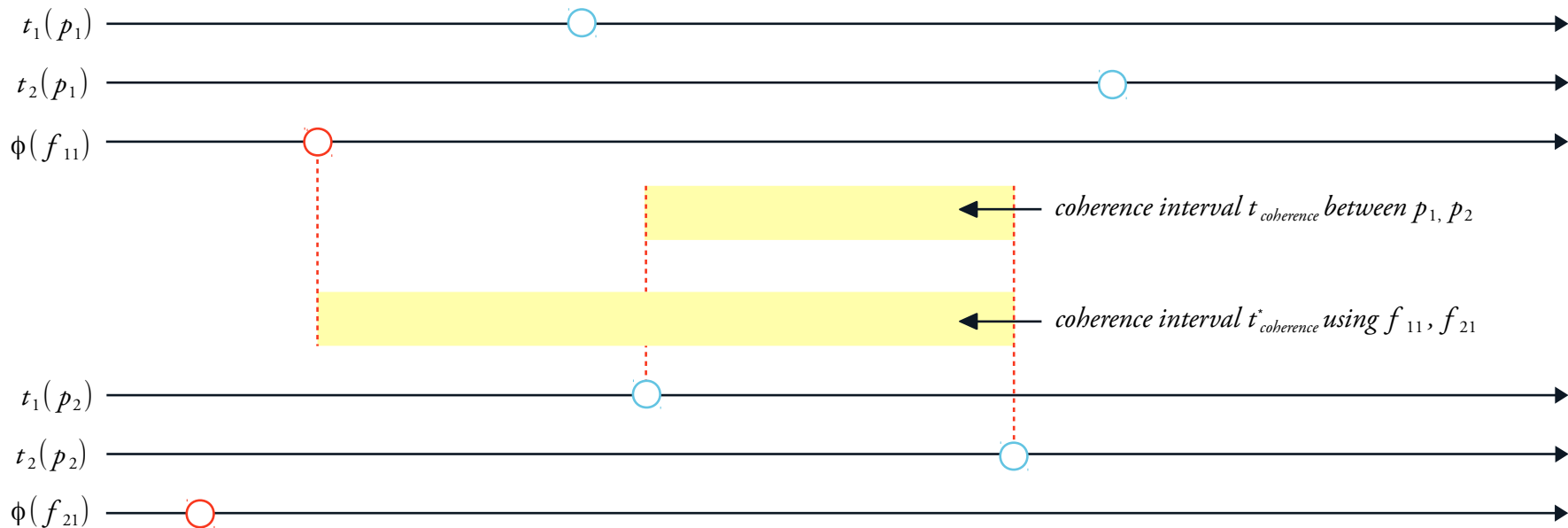
For yabiladi.com quartiles of $t_i(p_j) - \phi(f_{jk})$ in days are : (Q1) 256, (Q2) 777, (Q3) 1340

Upscaling the exploration (2/3)

> Disaggregated observable coherence

We define a discrete subset of fragments of interest

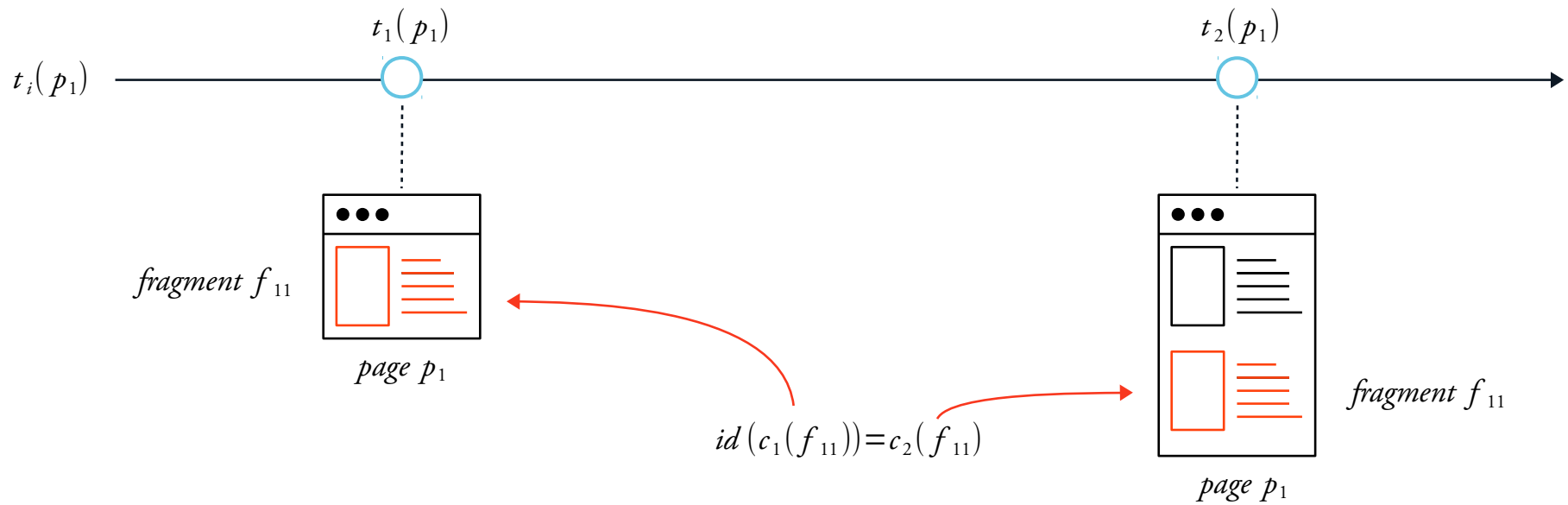
$$\forall p_j, \forall f_{jk}^* \in \{f_{j1}, \dots, f_{jm}\}, \exists t_{coherence}^*: t_{coherence}^* \in \bigcap_j [\phi(f_{jk}^*), t_i(p_j)] \neq \emptyset$$



And introduce a more permissive coherence model based on a specific research question

Upscaling the exploration (3/3)

> Duplicated archived contents

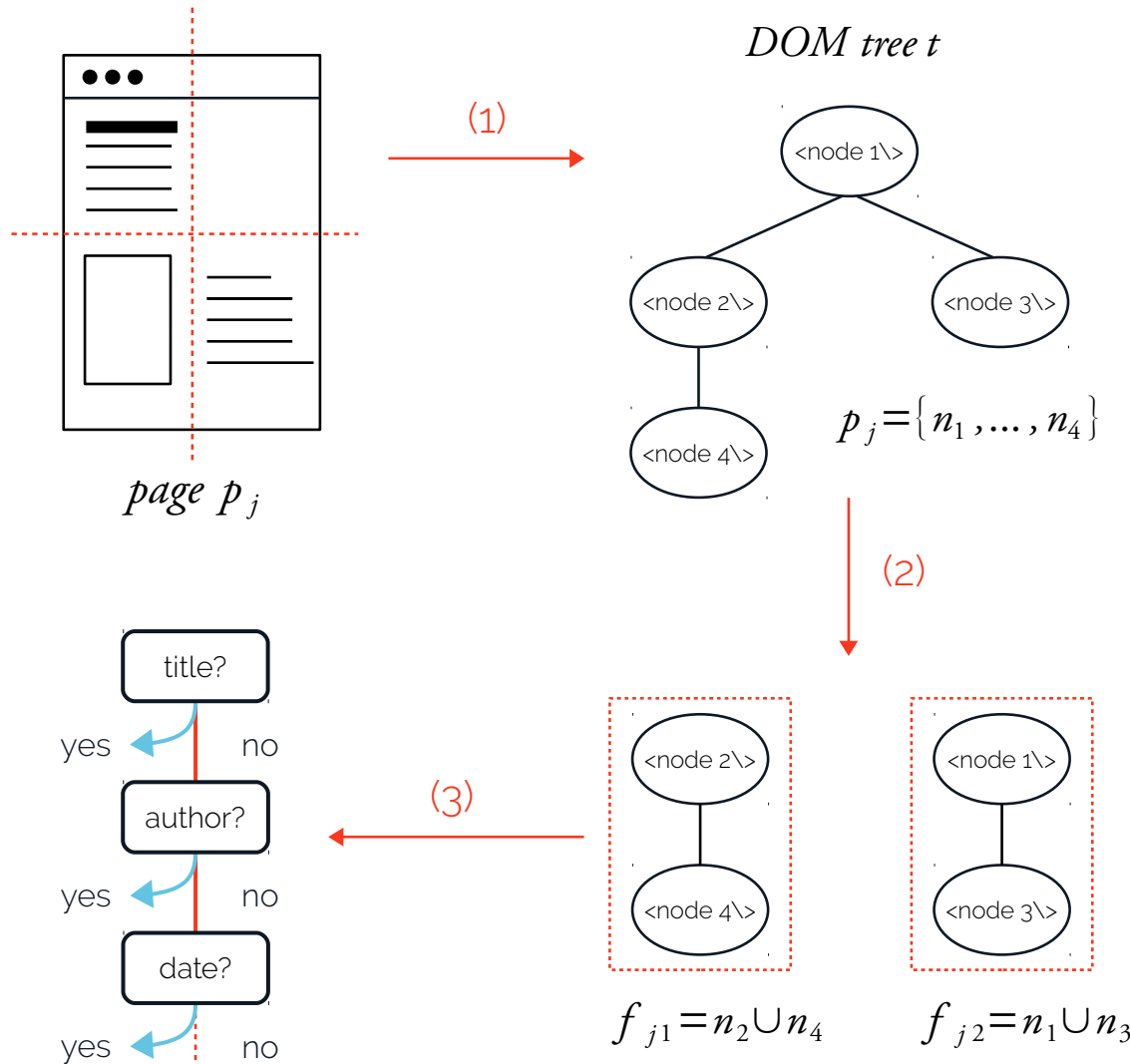


In practice, we deduplicate with a $id(sha256)$ on each Web fragment

For yabiladi.com quartiles of duplicated fragments : (Q1) 1, (Q2) 1, (Q3) 2, (Max) 44

Finding Web fragments

> Technical fragmentation and information extraction

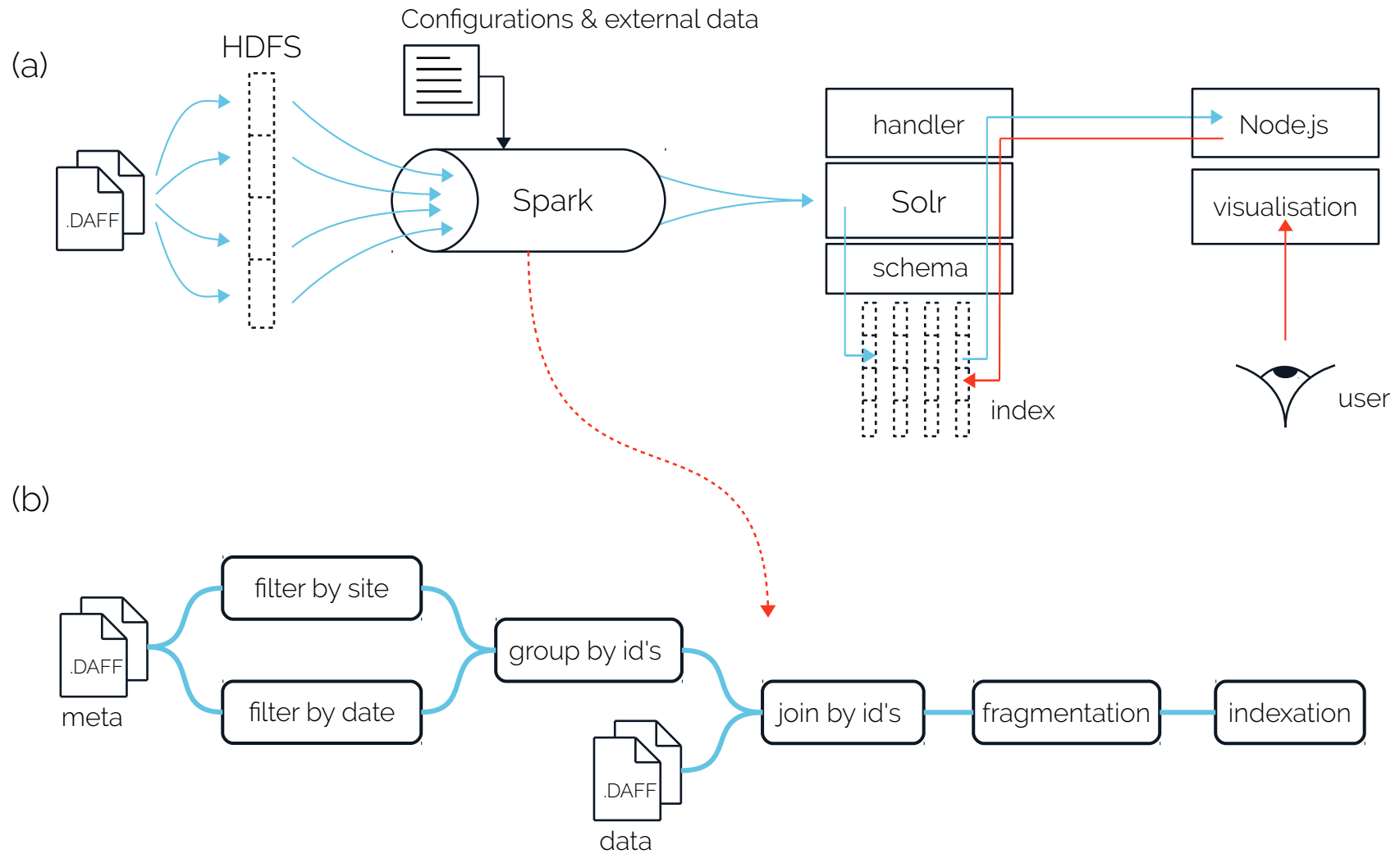


> Clustering closest HTML nodes using Readability and Fathom

> Distance function relies on vision / tag based penalties and ad-hoc rules. It can be set up **by the researcher**

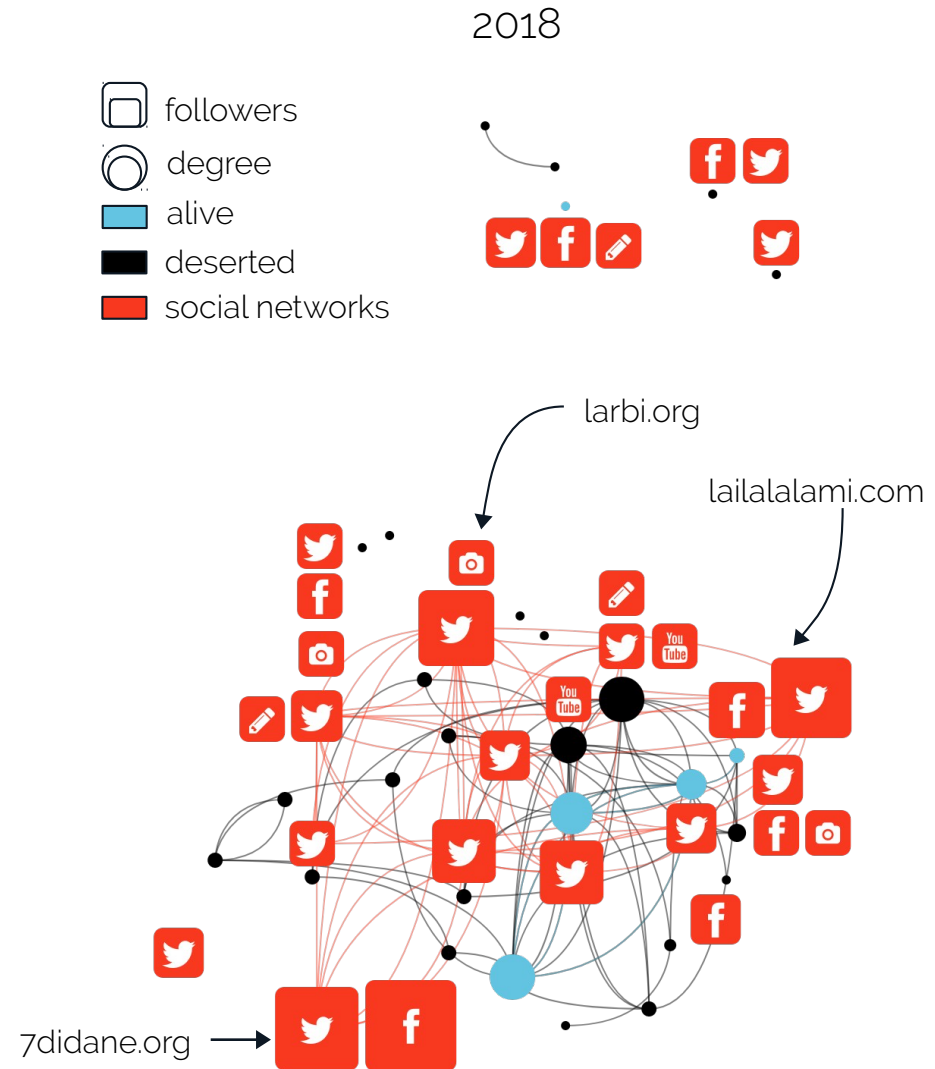
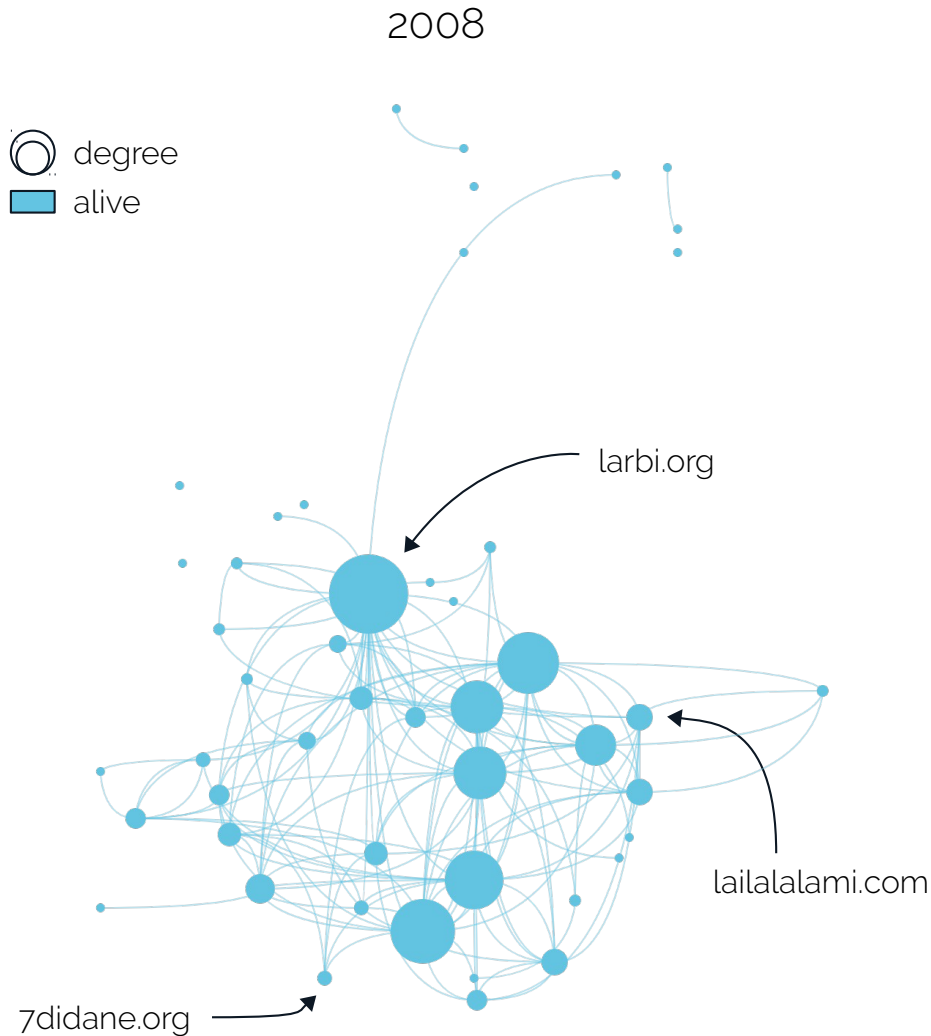
Building an exploration engine

> From archive files to search and visualisation facilities



The archived traces of digital mutation (1/3)

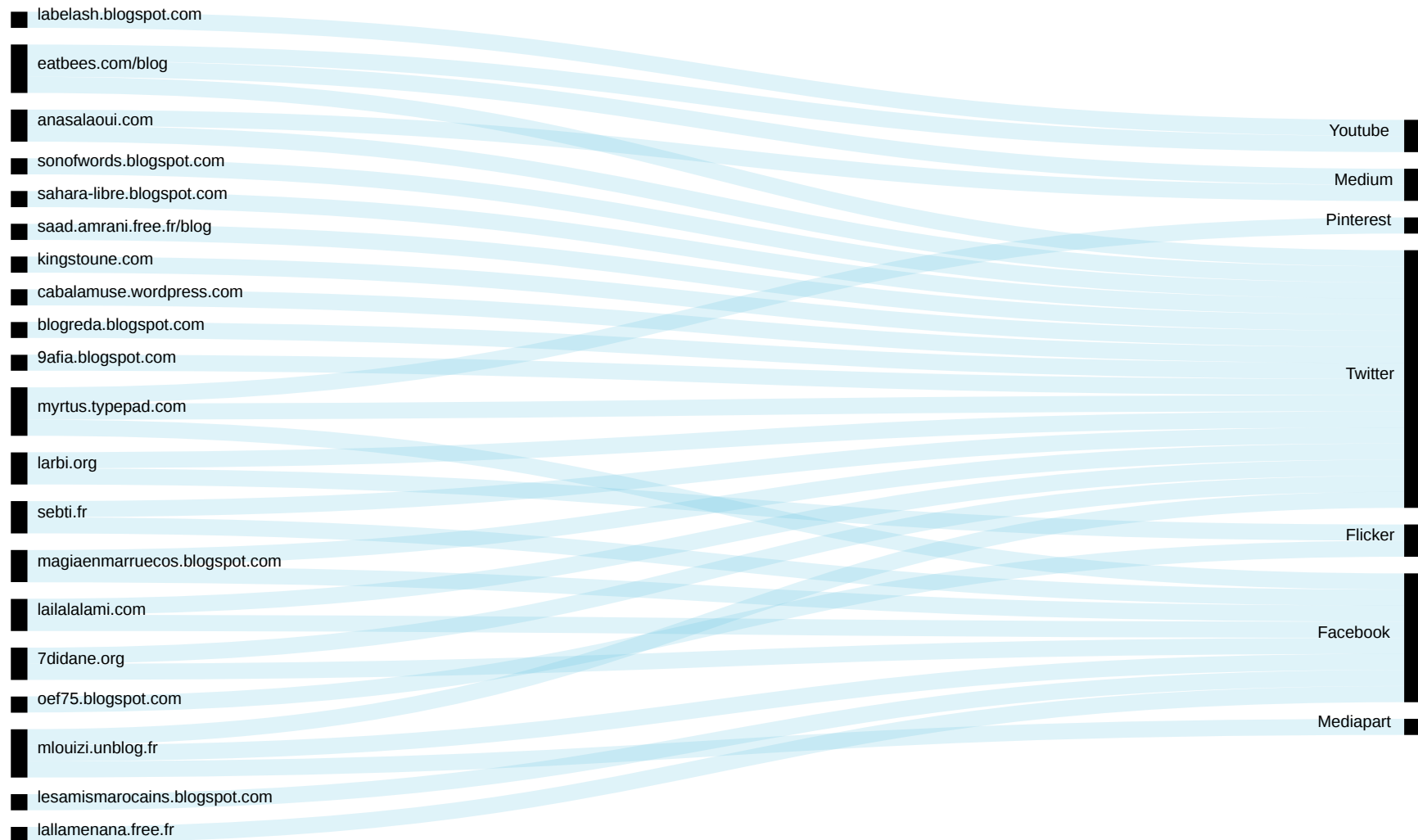
> Finding fragments mentioning social networks ``, Facebook



Authors kept their pseudonyms (or a close variation) from blogs to social platforms

The archived traces of **digital mutation** (2/3)

> Moving into new Web territories

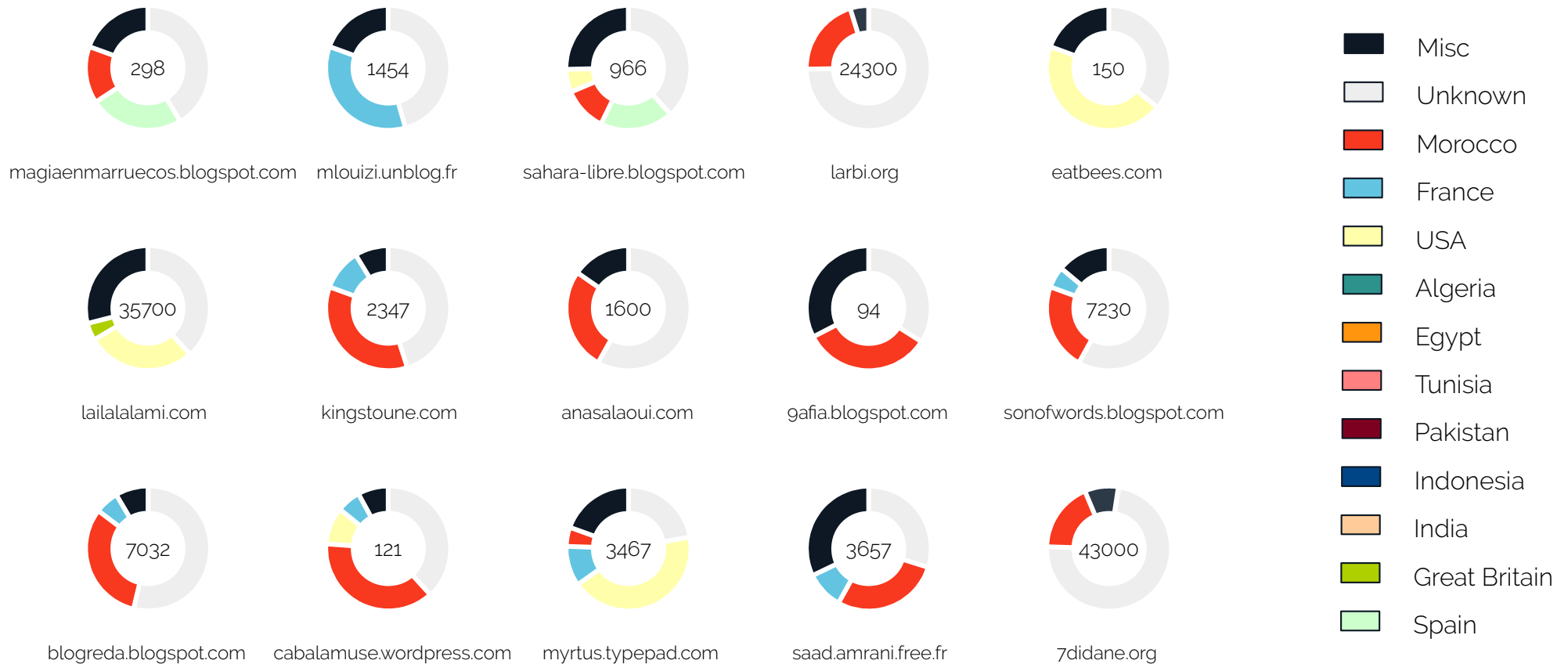


The expression is fragmented and specialized by type of medium

Graph density went from 0,16 in 2008 to 0,24 in 2018 (blogs vs twitter)

The archived traces of digital mutation (3/3)

> The recomposition of the community followed by the readers on Twitter



Readers followed larbi.org on Twitter
(26 % of the comments)

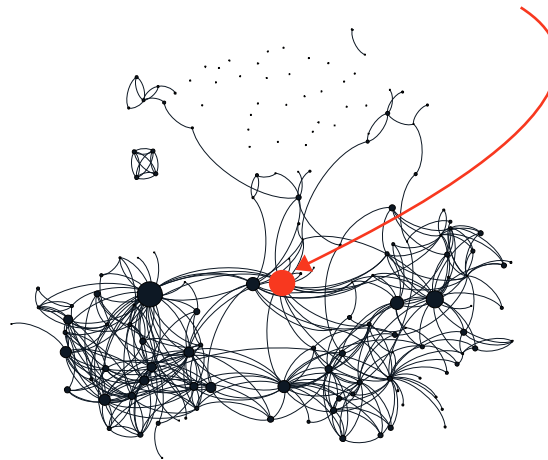


But the protest of February 20th 2011 (ash-tag #20Fev) seems to have played a key role in the mutation

"Morocco #Feb20 Maroc
Non le printemps arabe ne peut pas s'arrêter aux
Frontières du maroc – en direct de Twitter"

> larbi.org, 14 Feb 2011

> Does the M20F have influenced other part of the Moroccan e-Diasporas?
such as the old Web portal yabiladi.com ...



yabiladi.com

manual search

"20 février"

threads V_0

12 threads
94 users E_0

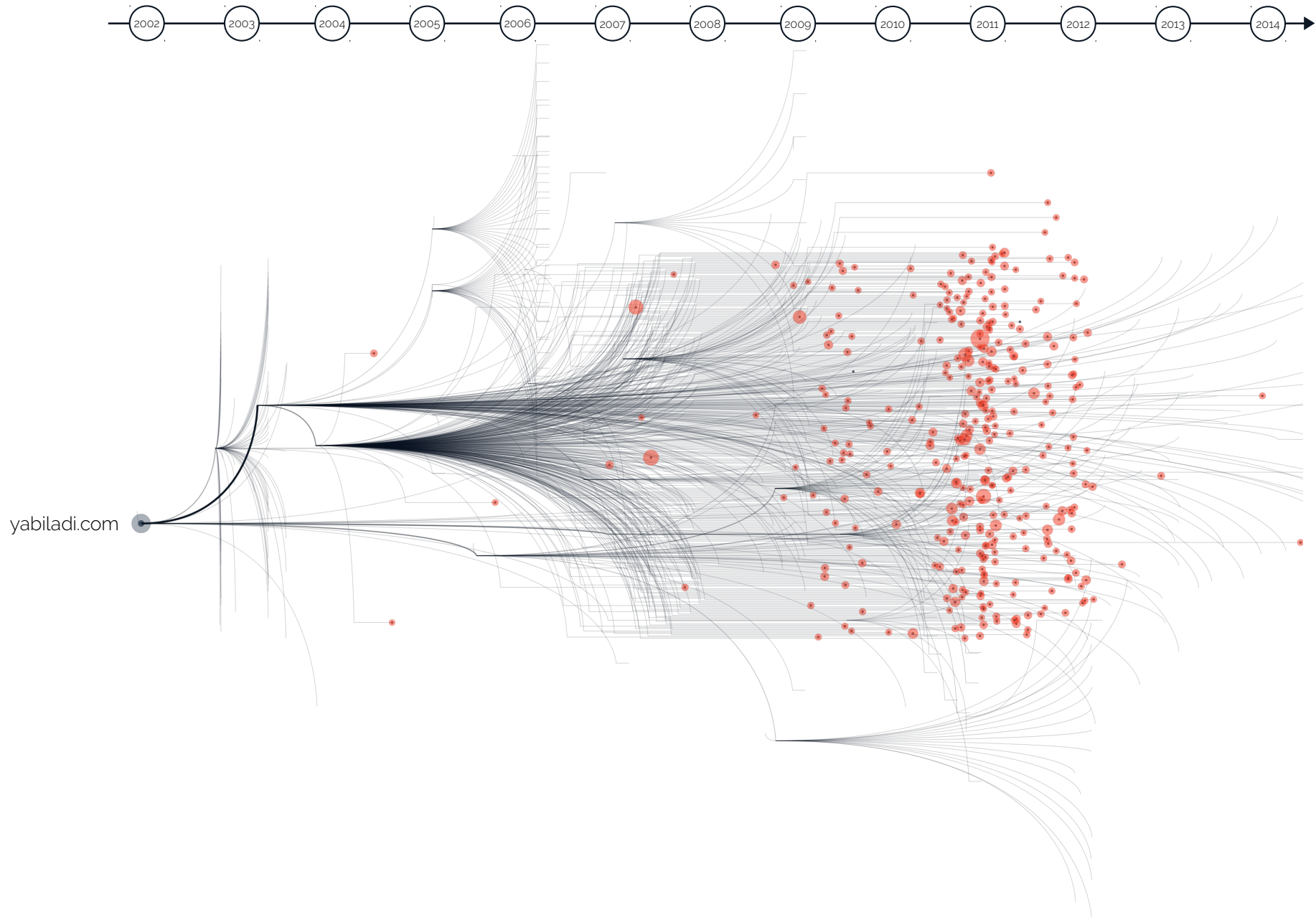
find co-contributors

threads V_1

341 threads
94 users E_0

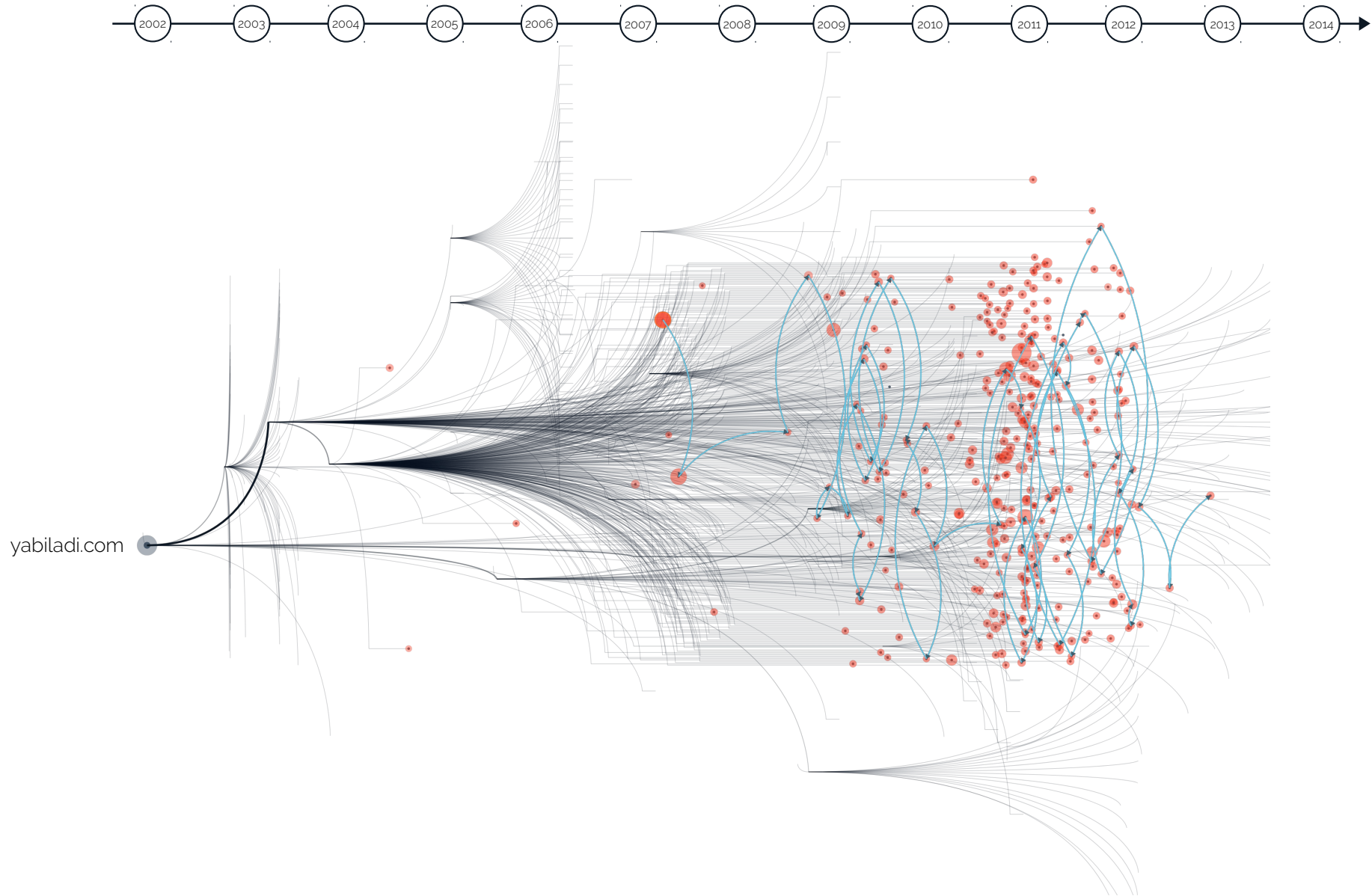
An **ephemeral** protest collective (1/4)

> Finding networks of relevant threads in yabiladi.com



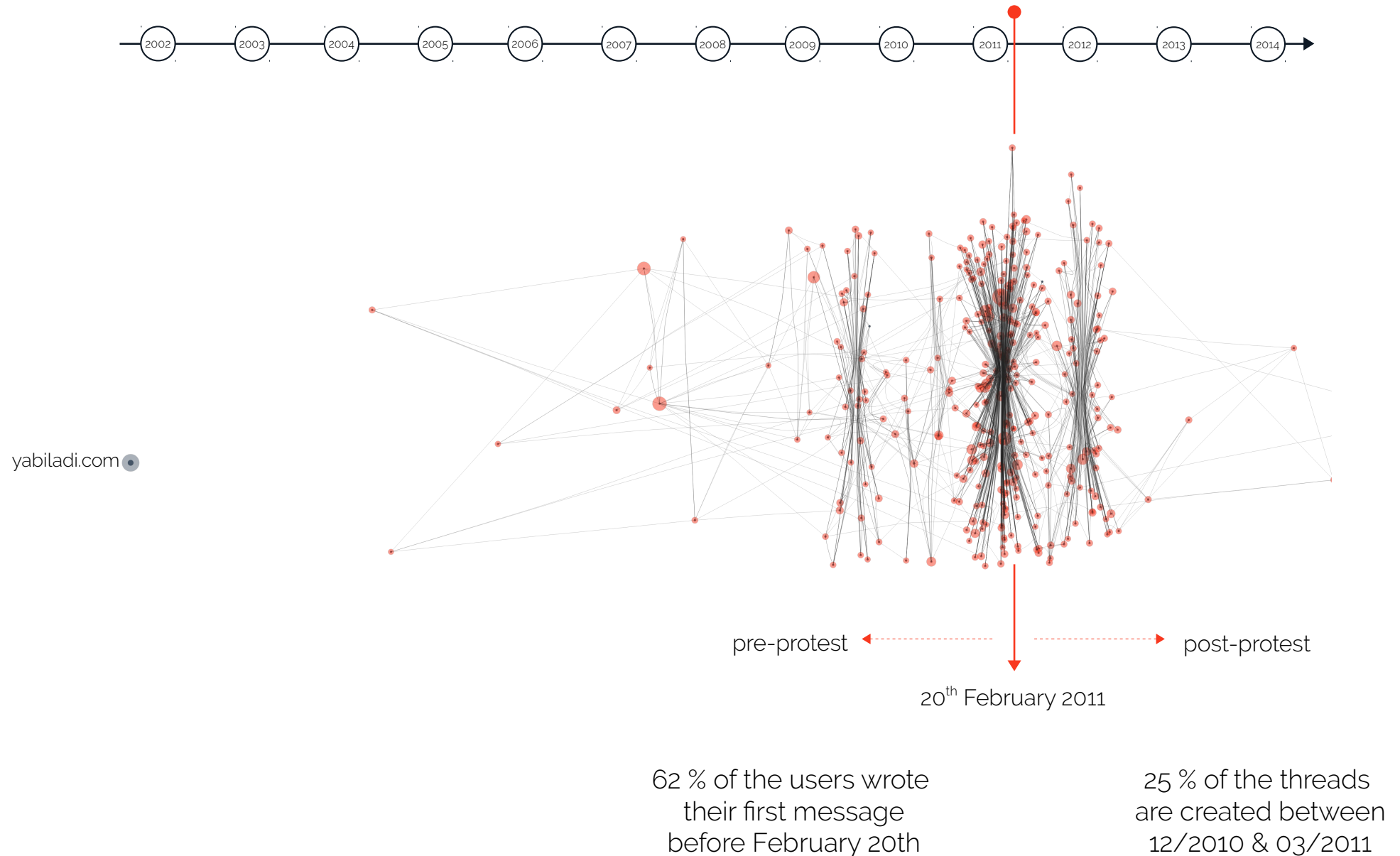
An **ephemeral** protest collective (2/4)

> Following users paths



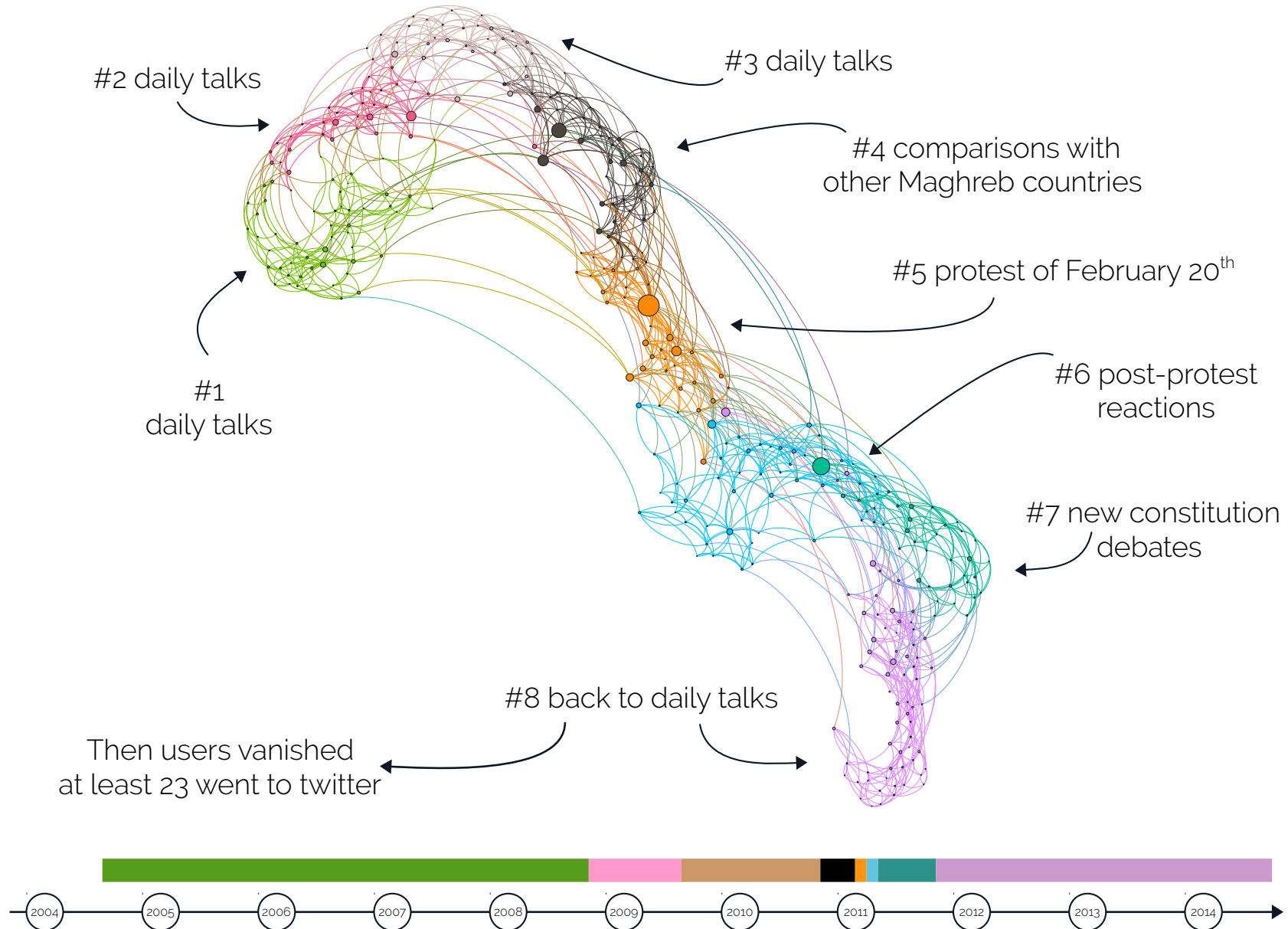
An **ephemeral** protest collective (3/4)

> Old members converge and new users directly join



An **ephemeral** protest collective (4/4)

> A sudden spark fires a minor part of the forum



But here we reach one of the limits of Web archives corpora and should consider the idea that Web archives may be intrinsically incomplete.

Web archives corpora only witness
the first leap of what we call a **pivot moment of the Web**.

Implication for **historical Web** studies

> Pivot moment of the Web

Web archives corpora still fail to convey the web as an ecosystem. While we were looking at the archived consequences of Arab Spring, Web actors were already moving away from forums and blogs.

In the same way as the long history of writing that was punctuated by key moments, the Web and the Internet in general already possess their own micro-history.


> We call **pivot moment of the Web** a period of transition between two systems, a moment when new Web uses fork from established habits and create gaps. A pivot moment arise from three factors: the **convergence** at **a specific moment** between **a technological leap** and a group of **users sieving it**.

Thank you !
Questions?

You want to go deeper into
Web archives and digital diaspora?



Good news !



My Phd's defence will take
place the 9th of November at
14:00 in amphi emeraude (B217)
there will be home made jam
and home brewed beer !

Quentin Lobbé (LTCI, Télécom ParisTech, Université Paris Saclay & Inria)
quentin.lobbe@gmail.com