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 e-Diasporas Atlas (1/2)

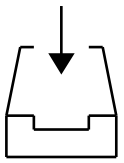
> A collection of online migrant collectives

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

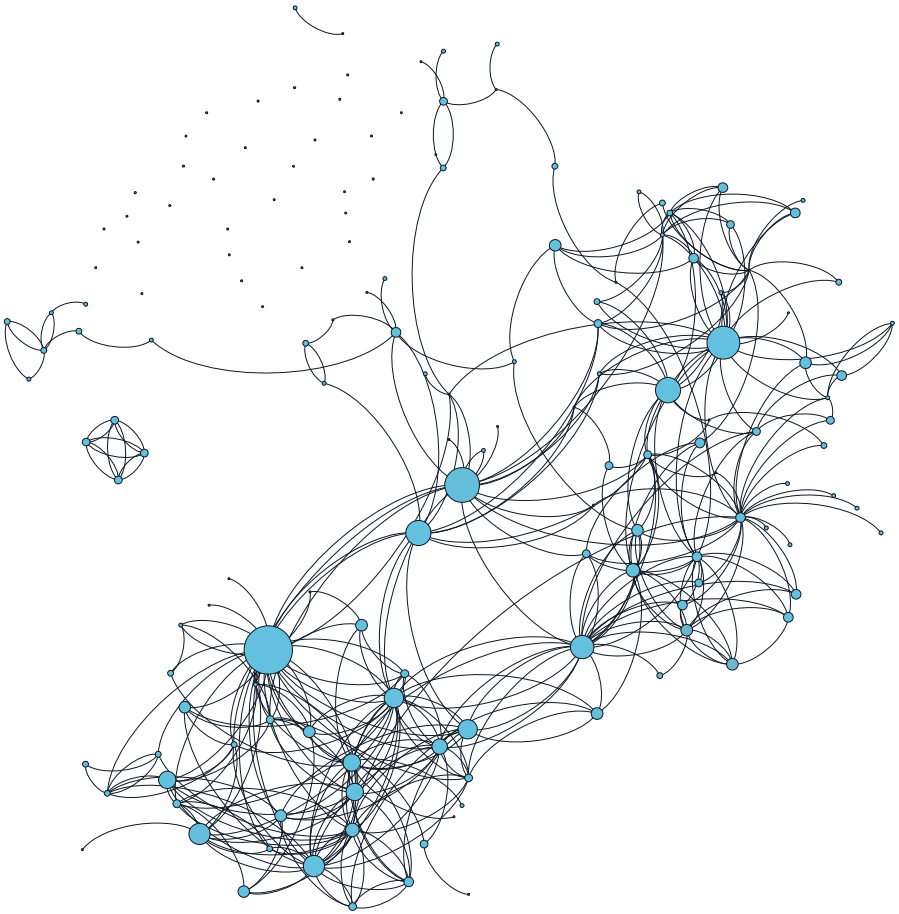
An e-Diaspora is a directed network of migrant Web sites linked by url (href)

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

> A corpus of Web archives



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

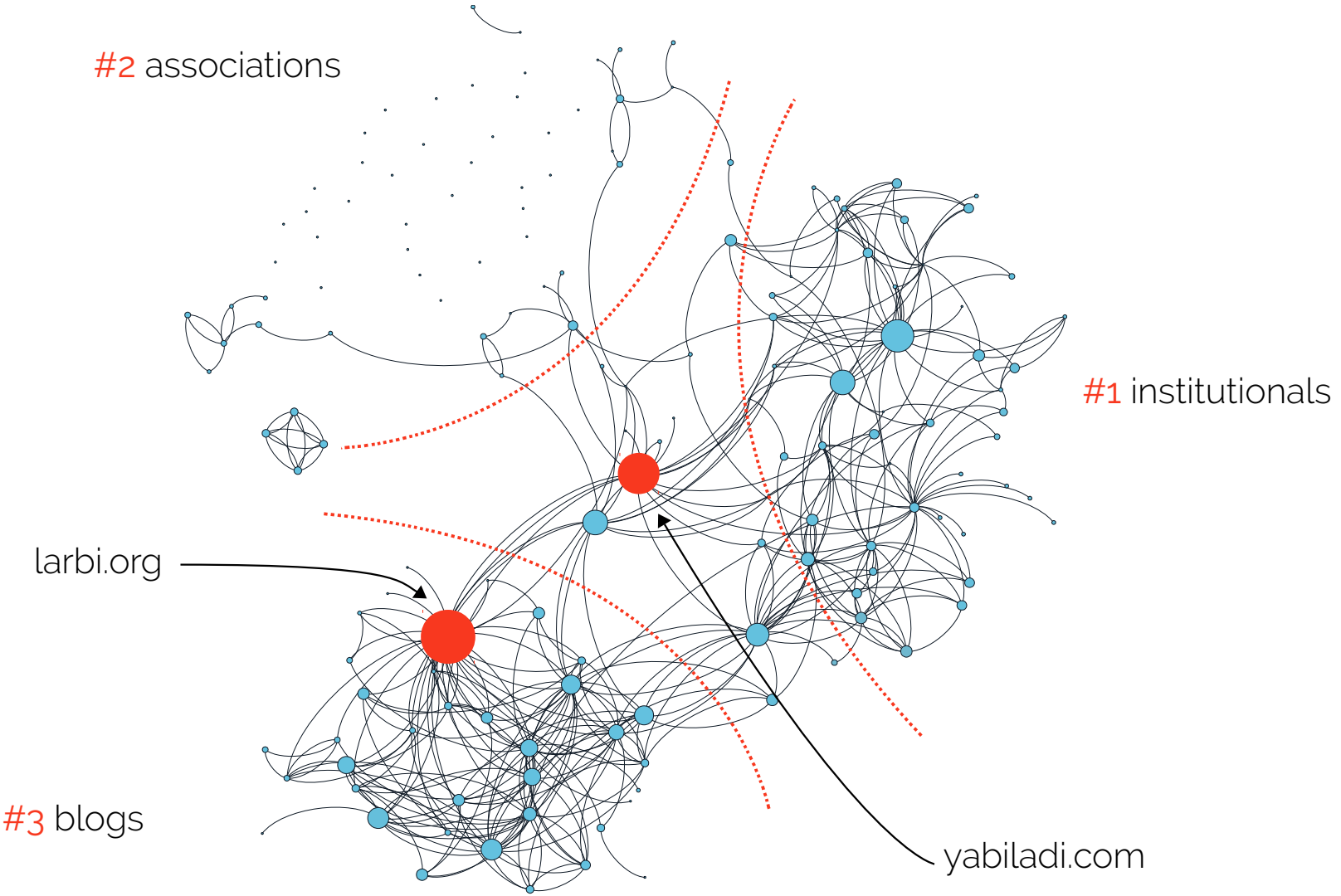


> See http://e-diasporas.fr/ (Dir. Dana Diminescu)

> The Moroccan e-Diaspora, by Dana Diminescu & Matthieu Renault

The e-Diasporas Atlas (2/2)

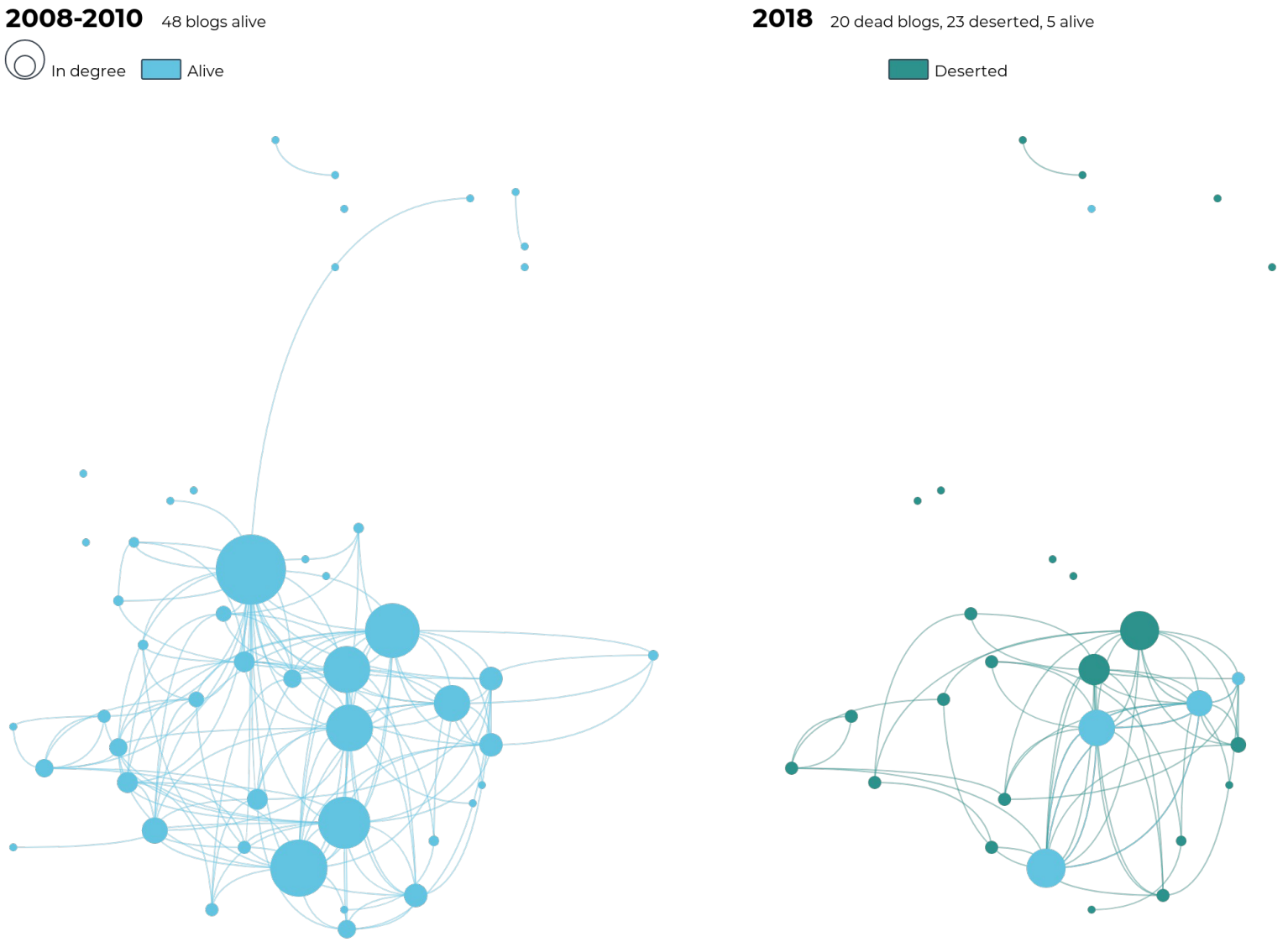
> Reading the map



> The Moroccan e-Diaspora, by Dana Diminescu & Matthieu Renault

An extinct online collective

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



> The Moroccan blogosphere (close up and evolution)

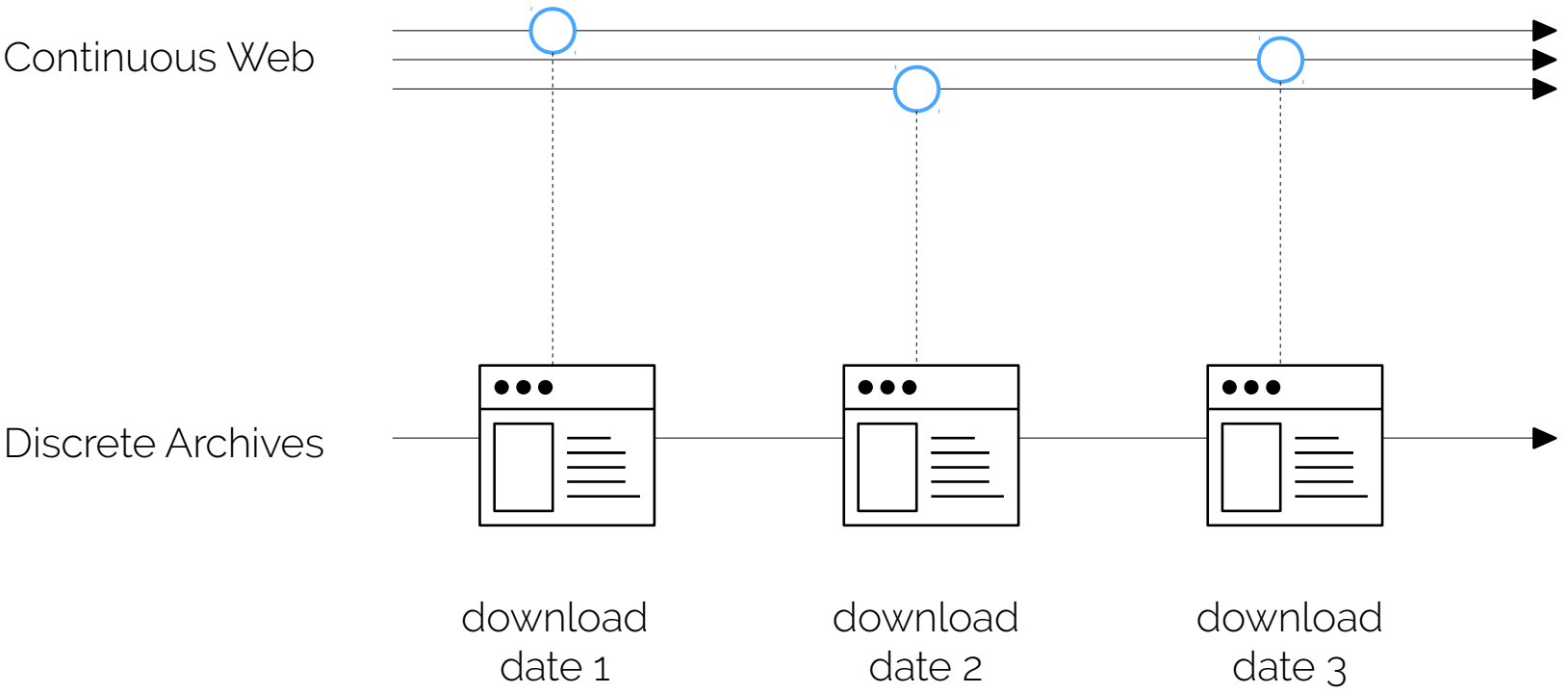
What happened to the dead Moroccan blogs?

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-Disaporas corpus of Web archives to find their remaining archived traces.

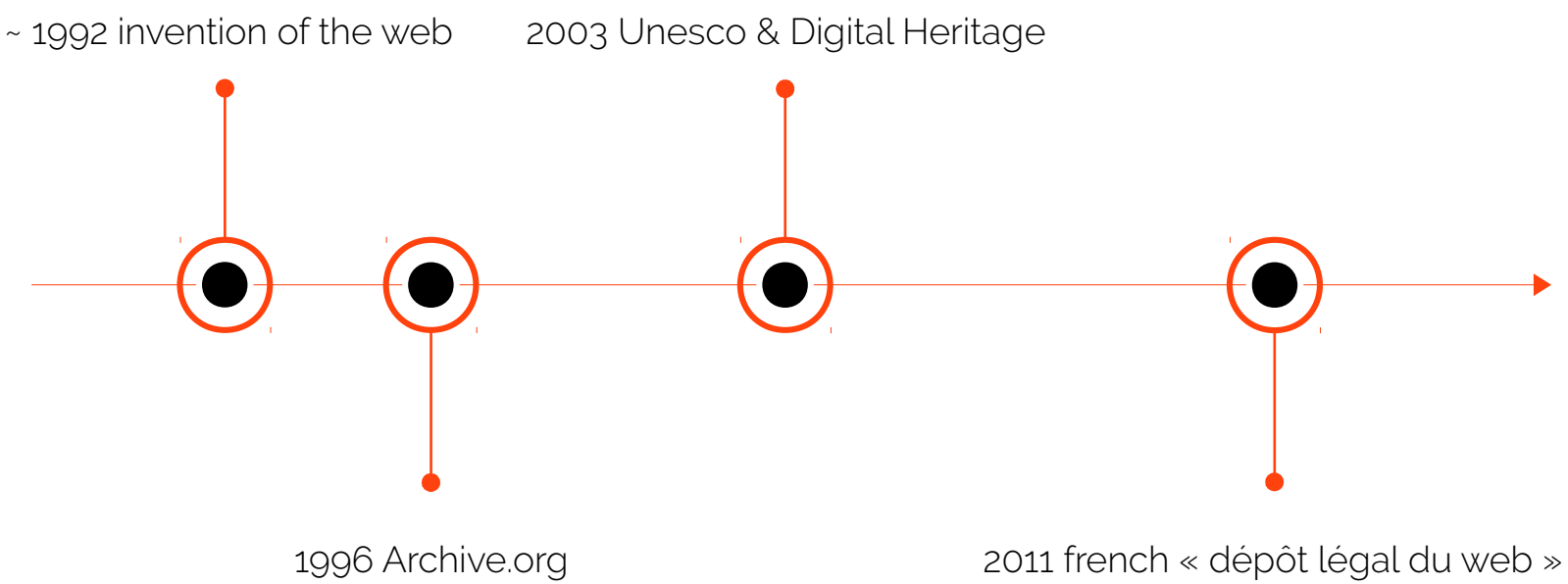
20 years of web archiving (1/2)

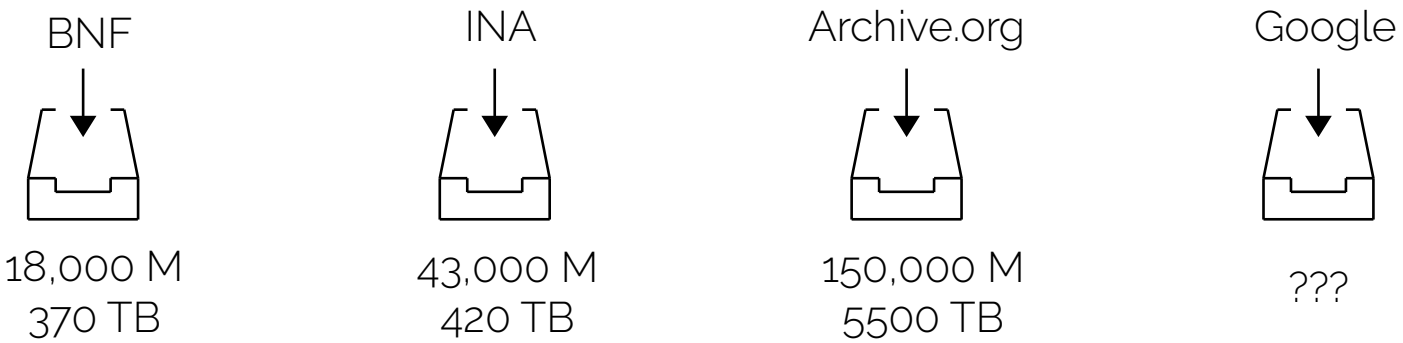
> Archiving our digital heritage



20 years of web archiving (2/2)

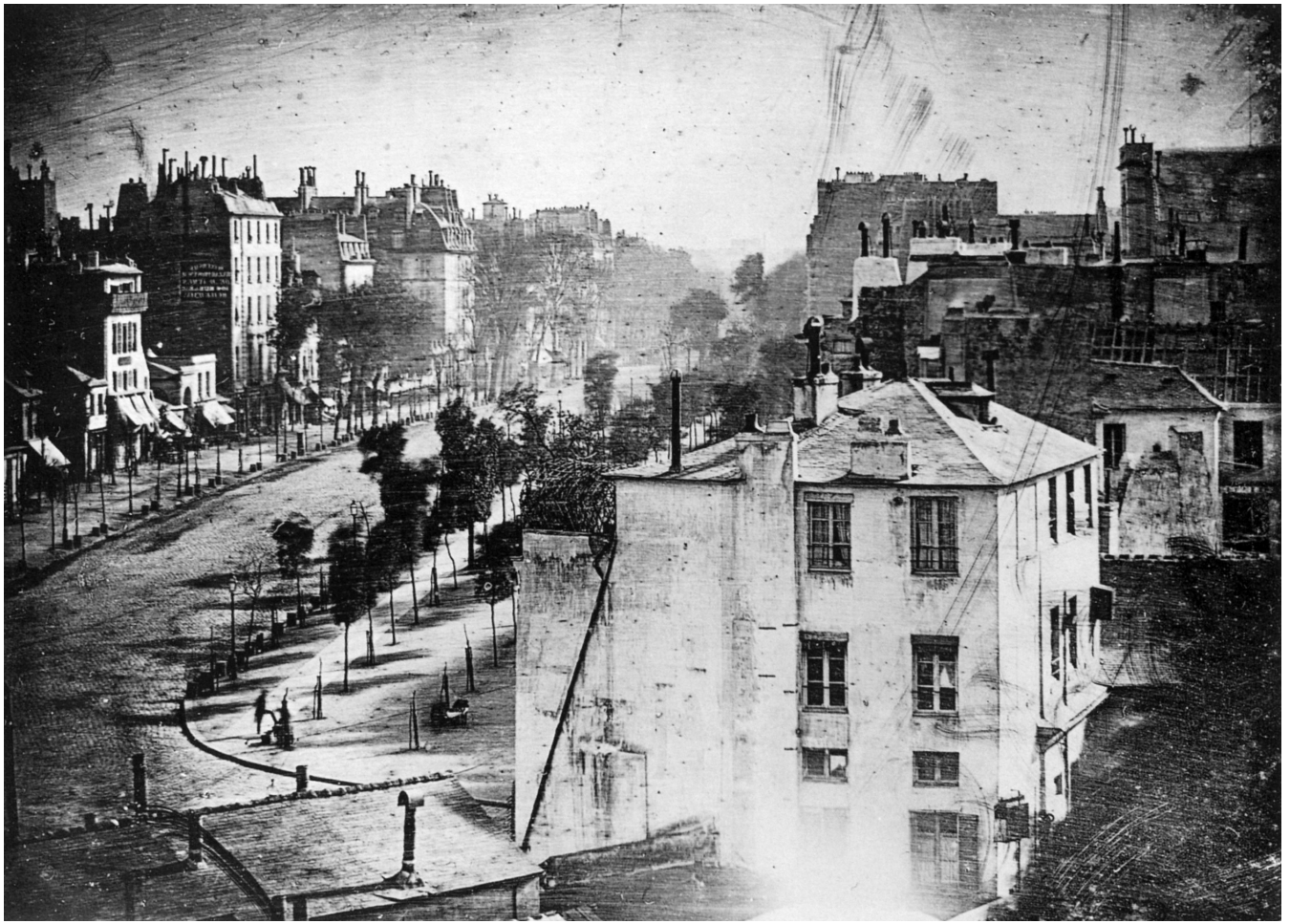
> Archiving our digital heritage





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

> Web archives are direct traces of the crawler

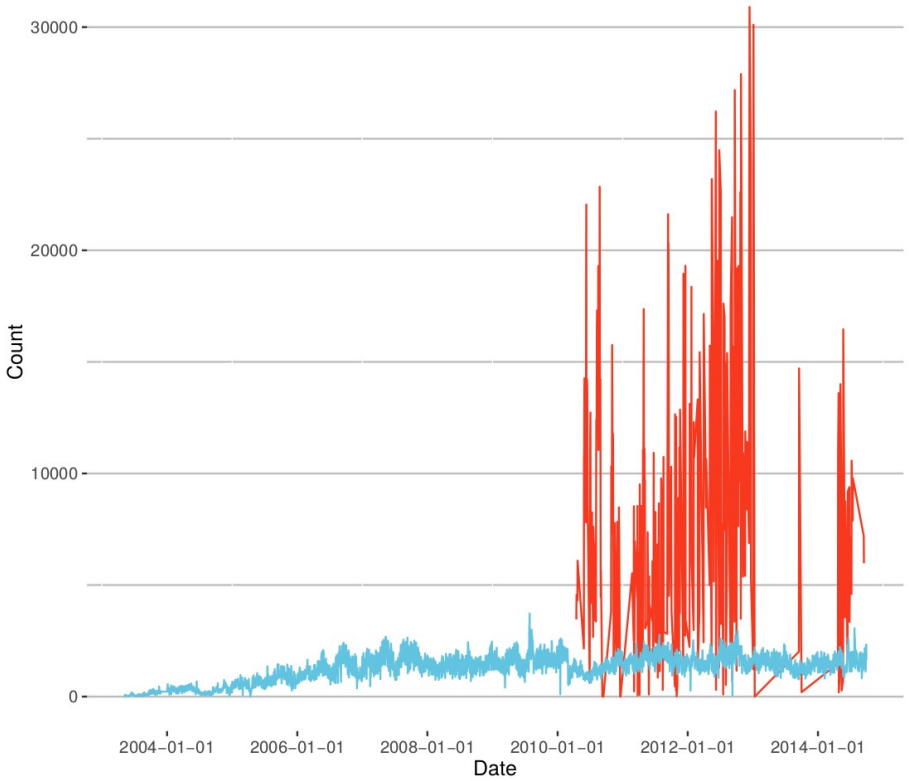


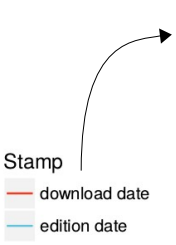
> "Boulevard du Temple", Louis Daguerre, 1838

By construction (WARC or DAFF files), 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





given by the archives file

to be found in the archives

> Distribution of the archived pages of yabiladi.com using download dates versus edition dates

We propose to introduce a new unit of exploration of Web archives corpora
to avoid all king of crawl legacy effects and maximise
the historical accuracy of an exploration

The Web fragment (1/3)

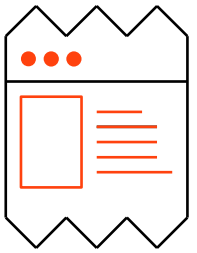
> Definition

Considering the Web page as the unit of access and consulation 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.

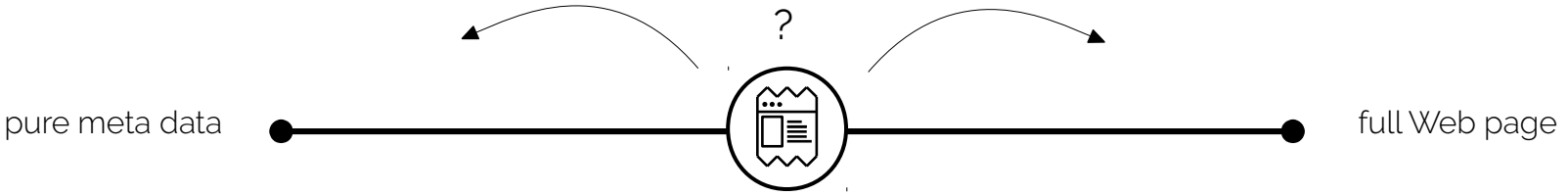
The Web fragment (2/3)

> Definition

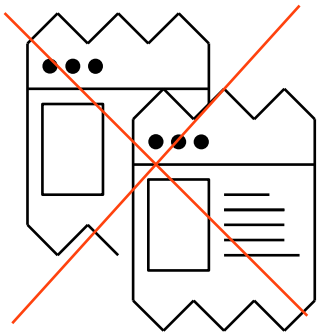
It's a coherent set of textual, visual or audio content that can be understood on it's own



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



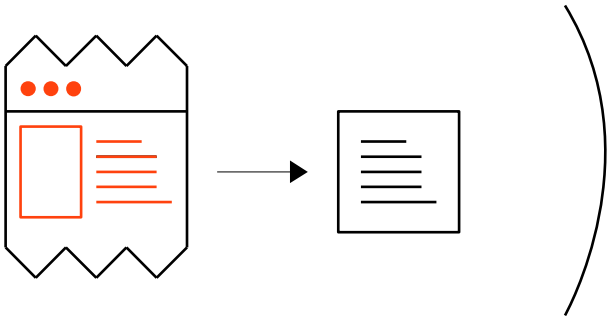
Within the same Web page, two Web fragments cannot overlap



The Web fragment (3/3)

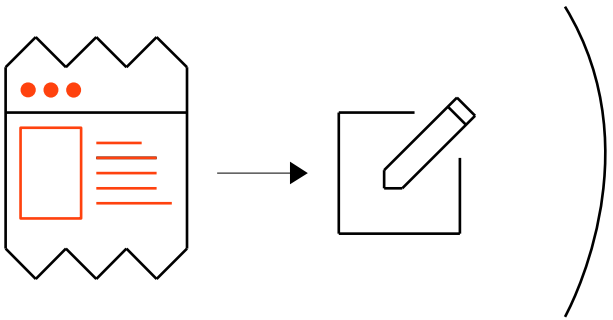
> Definition

It goes with an associated set of extracted meta contents



Title
Author
Date (edition date)

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

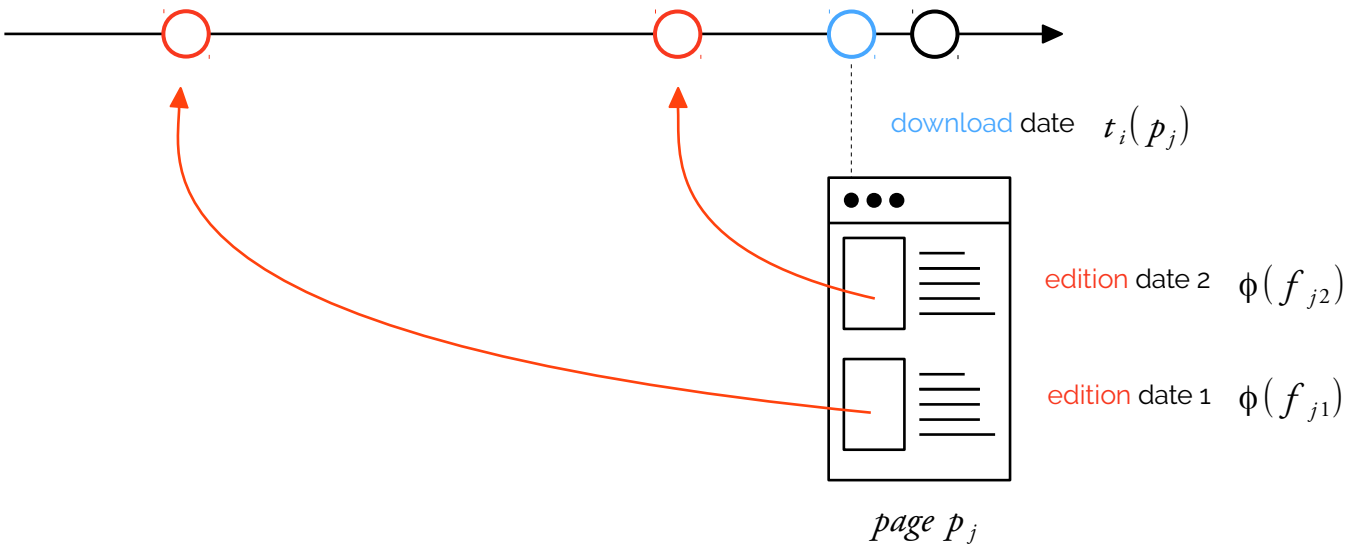


CMS widgets
Integrated text editor
Href links
Rss feed

Upscalling 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

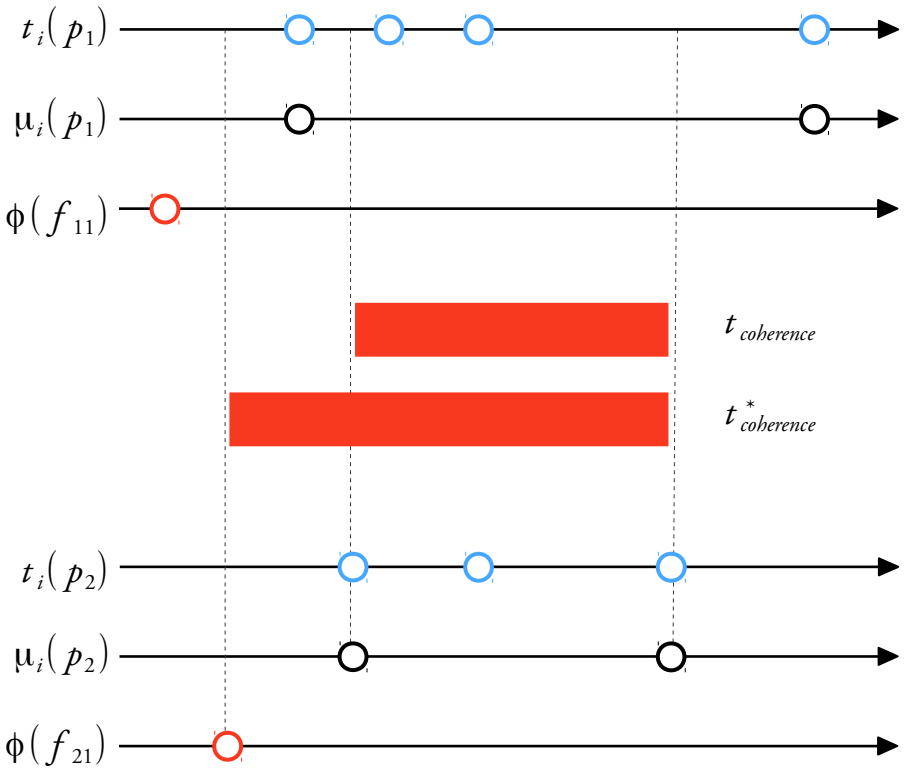
Upscalling the exploration (2/3)

> Disaggregated observable coherence

We define a discret subset of fragments of interest so that :

$$\forall\, p_j, f^*_{jk} \in \{f_{j1}, \ldots, f_{jl}\} \exists\, t^*_{coherence} :$$

$$t^*_{coherence} \in \bigcap_{i=1}^{n} [\phi(f_{jk}), t_i(p_j)] \neq \emptyset$$

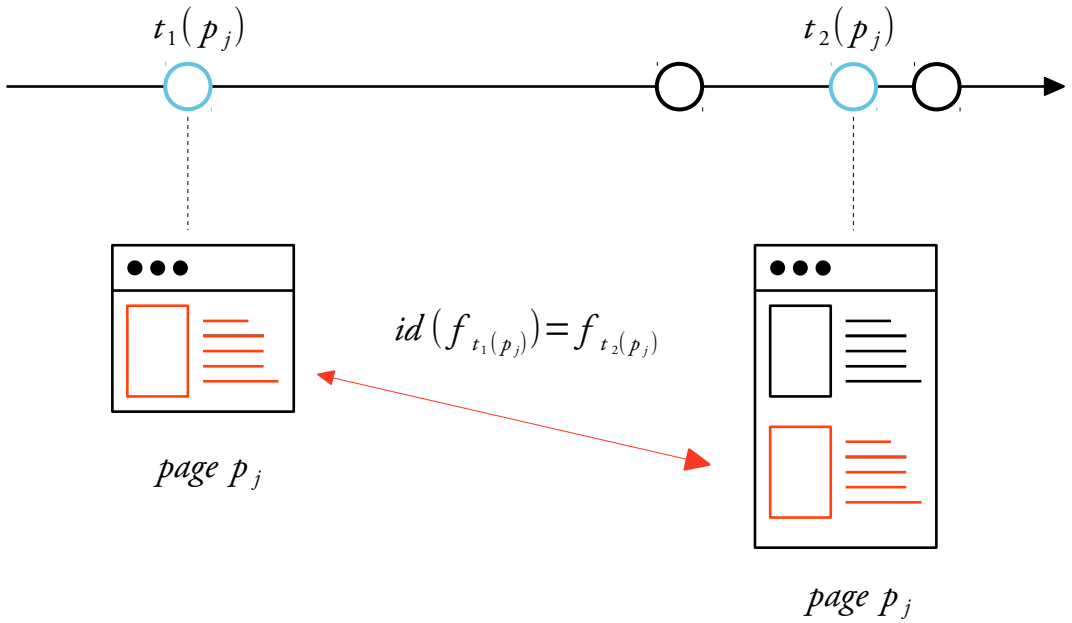


We introduce a more permissive archive coherence based on a specific research question

> M Spaniol. 2009. Data Quality in Web Archiving. (WICOW '09)

Upscalling the exploration (3/3)

> Duplicate archived contents



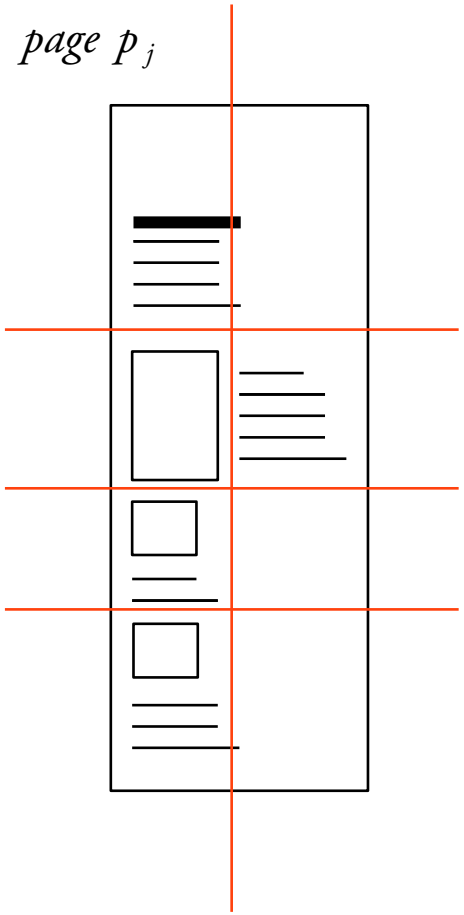
In practice, we deduplicate with a id(sha256) on each Web fragment

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

The level of indexation in the following, will be the Web fragment

Finding Web fragment

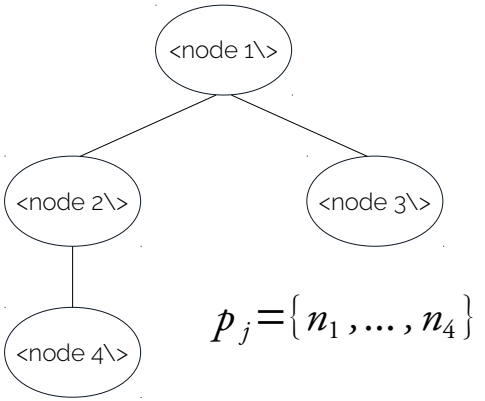
> Technical segmentation and extraction



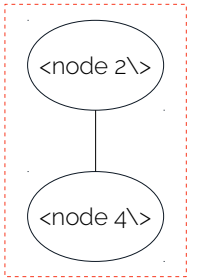
We clean the DOM tree
using boilerplate method

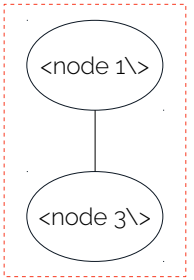
Readability and Fathom
methods for clustering closest
HTML nodes in an
adjancy matrix

Distance between 2 nodes
relies on vision based and tag
based penalties
we also add ad-hoc rules



$p_j = \{n_1, \ldots, n_4\}$





$f_{j1} = n_2 \cup n_4$

$f_{j2} = n_1 \cup n_3$

> See https://github.com/mozilla/readability

> See https://github.com/mozilla/fathom

> C. Kohlschütter et al, 2010. Boilerplate detection Using Shallow Text Features. (WSDM '10)

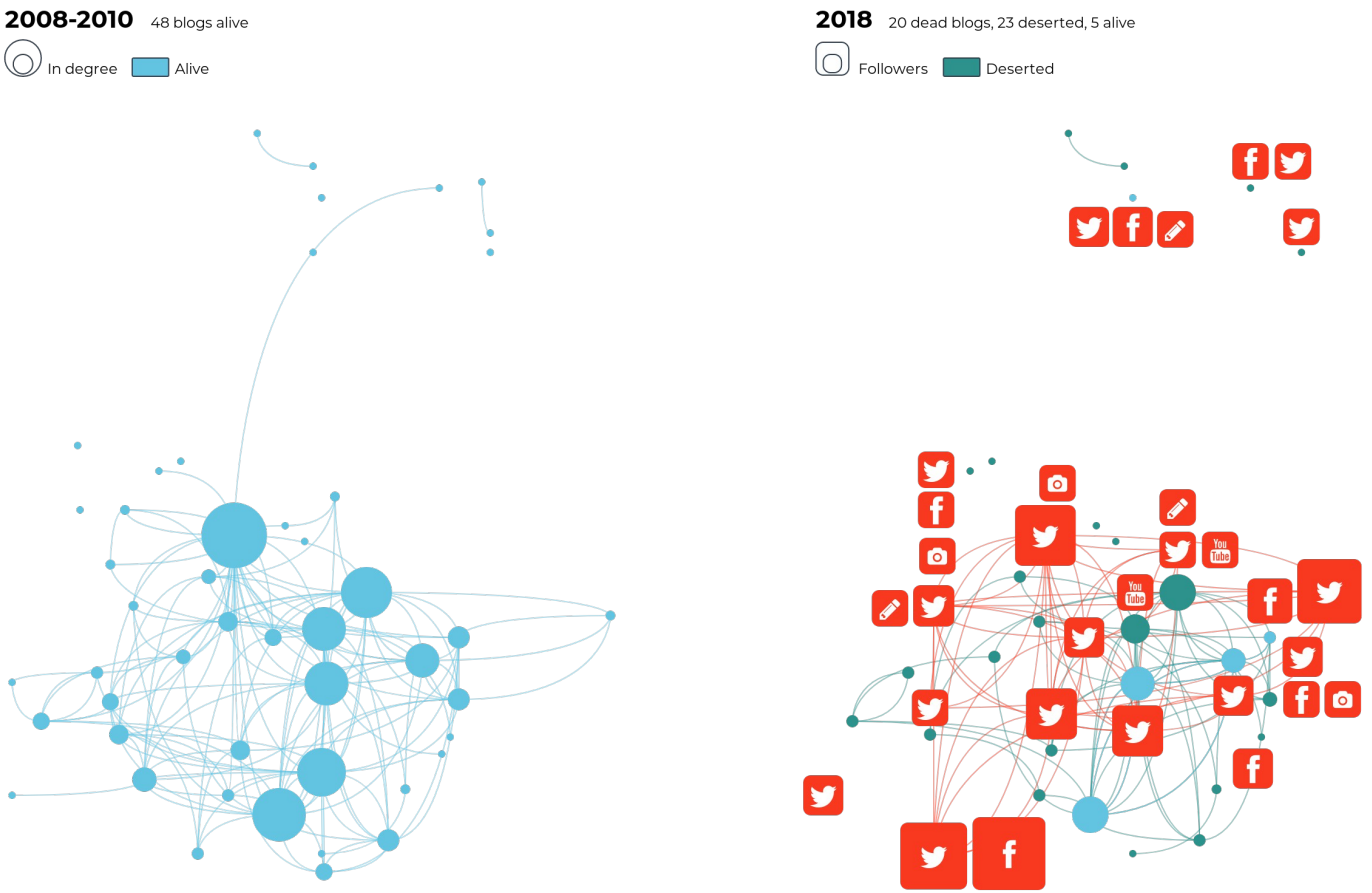
> D. Cai et al, 2003. Vips: a vision-based page segmentation algorithm. (2003)

> A. Jatowt et al, 2007. Detecting age of page content. (2007).

The archived traces of digital mutation (1/2)

> Exploration task

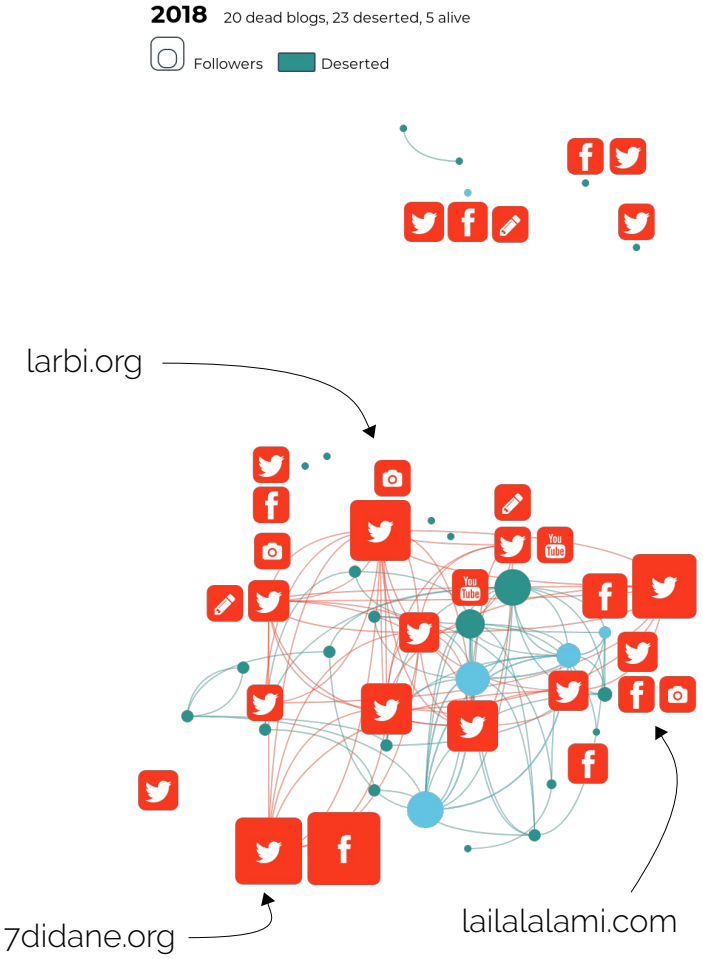
We request for ** or *<button class="Facebooks-share"></button>*
or Web fragments directly mentioning facebook, youtube, pinterest ...



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

The archived traces of digital mutation (2/2)

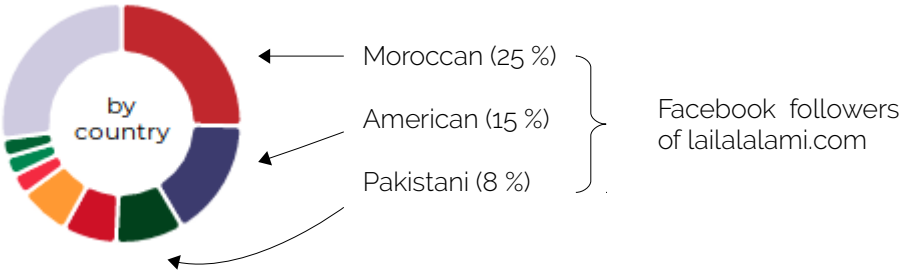
> The recomposition of the community followed by the readers



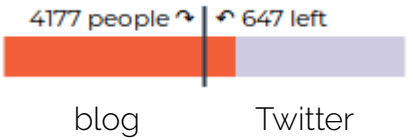
The expression is fragmented and specialized by type of medium

Graph density went from 0,16 in 2008 to 0,24 in 2018

Authors kept their diasporic readers



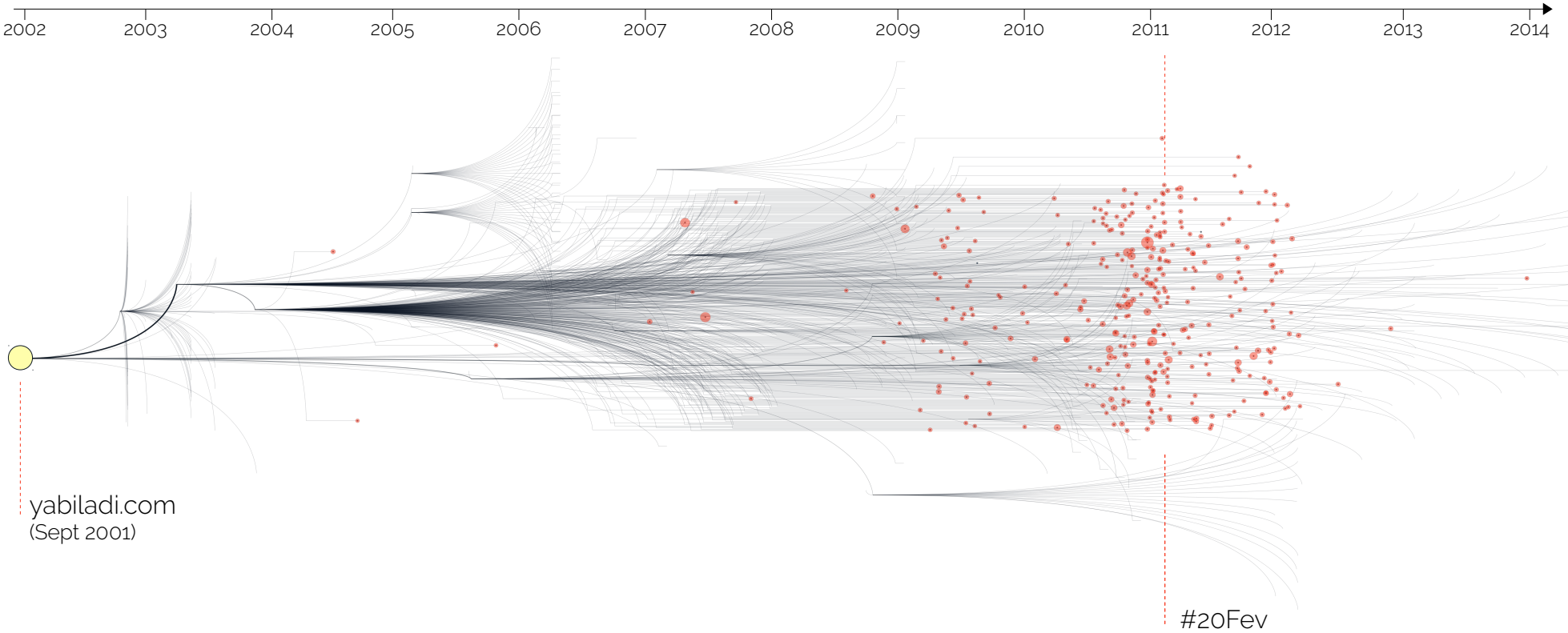
Readers followed larbi.org on Twitter (26 % of the comments)



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

An ephemeral protest collective (1/4)

> An old forum and a hub for Moroccan migrants : yabiladi.com



2,683,928 archived Web pages in e-Diasporas

We start with 12 topics ● matching «*20 février*», consisting of 196 messages written by a set of 94 unique users

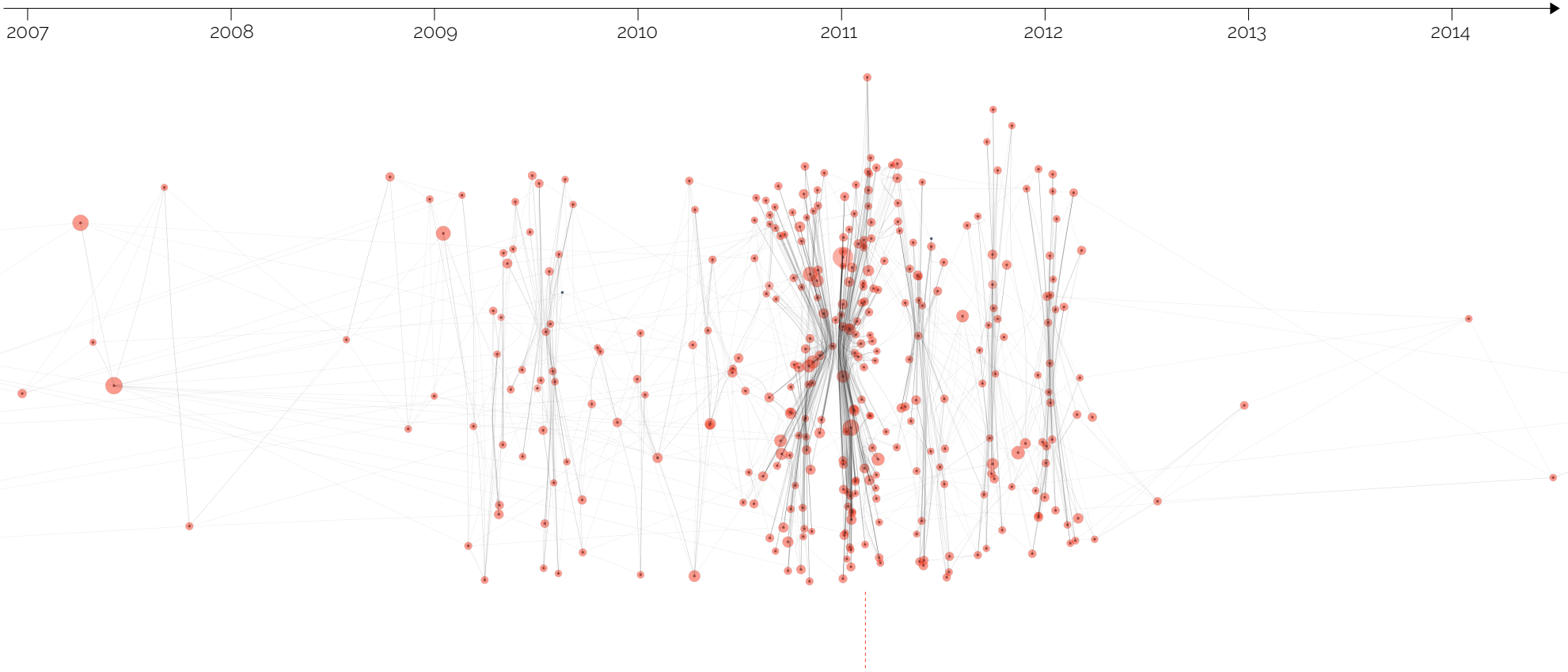
We select all the topics where at least 2 of those users wrote a message. This results in a network of 343 topics linked by co-contributors.

An ephemeral protest collective (2/4)

> Following users paths

An ephemeral protest collective (3/4)

> Old members converge and new users directly join



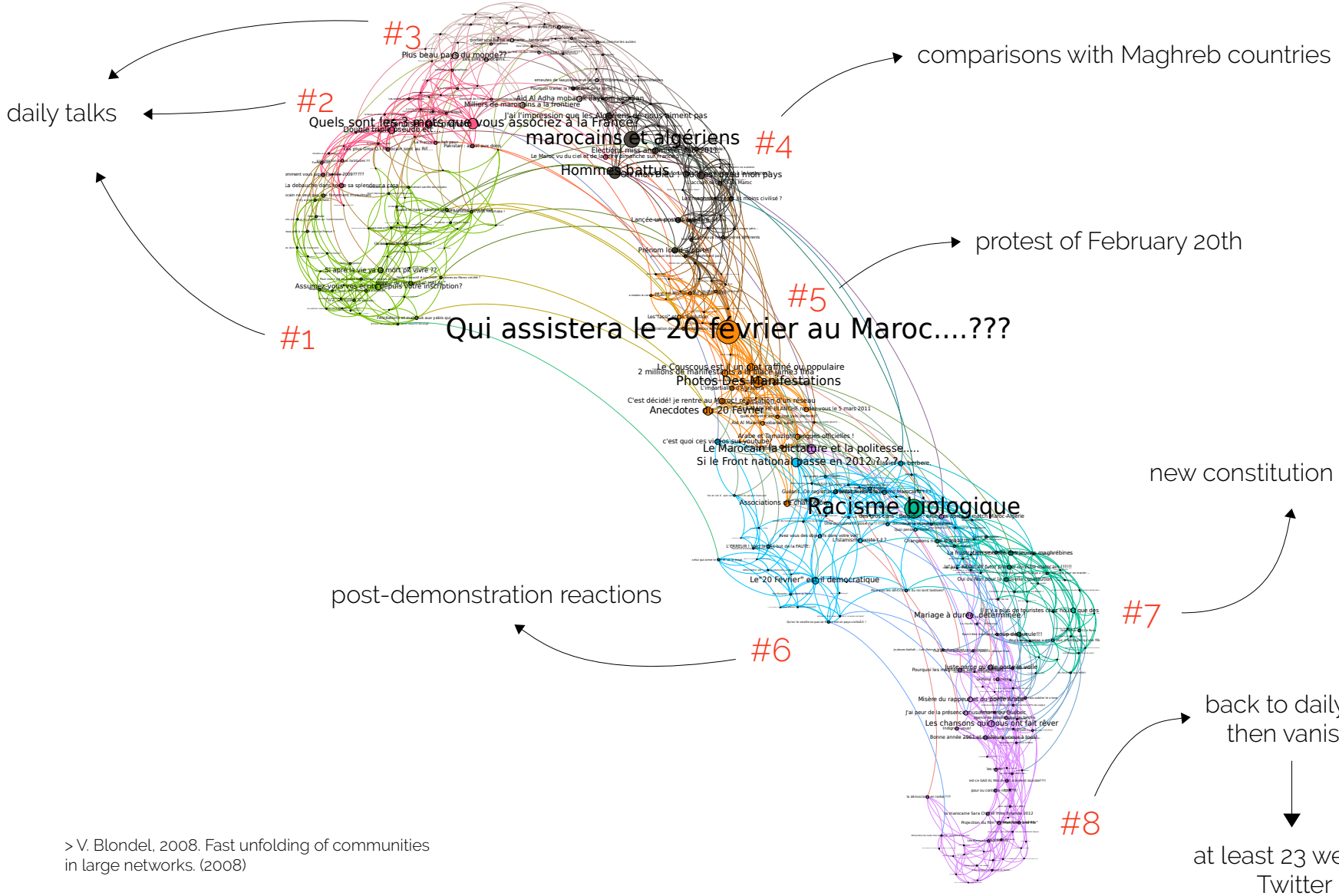
62 % of users wrote their first message before February 20th

#20Fev

25 % of topics are created between 12/2010 & 03/2011

An ephemeral protest collective (4/4)

> A sudden spark fires a minor part of the forum



daily talks

comparisons with Maghreb countries

protest of February 20th

new constitution debates

post-demonstration reactions

back to daily talks
then vanished

at least 23 went to
Twitter

> V. Blondel, 2008. Fast unfolding of communities in large networks. (2008)

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 possesse 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 users sieving it.

Thank you !
Questions ?

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