Continuity and discontinuity in web archives

Quentin Lobbé

Continuity and discontinuity in web archives

Quentin Lobbé[1*]

1 Complex Systems Institute of Paris Île-de-France, ISCPIF CNRS, Paris, France

Abstract

Web archival materials are not direct traces of the web, they are direct traces of crawlers. By design, the structure of web archives limits our collective capacity to explore the memory of the Web. These structural issues induce temporal discontinuities in the archives such as inconsistency, redundancy and blindness. In this paper, we address the question of re-injecting continuity within large corpora of web archives. We thus introduce the notions of persistences (series of time-stable snapshots of archived web pages) and continuity spaces (networks of time-consistent persistences). We demonstrate how – on the basis of a quality score – persistences can be used to select subsets of web archives within which in-depth historical analysis can be conducted at scale. We next propose to make use of a new visualization approach called the web cernes to graphically reconstruct the multi-level evolution of an archived web site. We finally apply our framework to study the archives of the *firsttuesday* movement: a constellation of networking web sites that acted in the interest of the economical growth of the web in the early 2000's.

1 Introduction

Memory institutions have been archiving the web for the last 25 years in order to preserve pieces of our digital heritage [1]. But just like *Funes* in Borges short story[1], national libraries and web archivists will never be able to achieve exhaustiveness. Indeed, the web is too wide, too complex and too profuse to be entirely preserved. Selection criteria stated by librarians, curators or politicians and applied by engineers and crawlers thus determine the spatio-temporal coverage of web archives corpora. But unlike traditional archival materials, the technical constraints of web archives influence our capacity to study their historical content. Web archives can't be understood apart from their own archiving processes: crawlers tear web resources away from the continuous temporality of the web[2] and produce discretized snapshots strictly timestamped by archiving dates. By design, '*web archives are not direct traces of the web, they are direct traces of crawlers*' [3].

1.1 How to unlock the historical richness of web archives?

The web preservation movement has originally been sparked with the intuition that web resources were intended to become valuable research materials in the hands of future historians; and archiving pioneers indisputably succeeded in preserving what could be

[1] See the fable of "*Funes the Memorious*" in Borges' anthology *Ficciones*.
[2] See [2] for a wide review of web archiving techniques.

saved in a short period of time[3]. Yet, the historical analysis of existing archives corpora remains an open problem. Convincing qualitative approaches do exist (see 1.3), but exploration tasks quickly become non-trivial when scholars widen the scope of their analysis. At a large scale (more than a hundred pages), collections of web archives suffer from being too broad or too rich and paradoxically incomplete: they lack spatio-temporal continuity.

In this article, we think that the key to unlock large scale analysis of web archives depends upon our capacity to resolve the problem of discontinuity induced by the technical constraints of such materials. We thus propose a theoretical and practical framework for reconstructing *persistences* and *continuity spaces* inside web archives corpora (see 2). Persistences are series of time-stable snapshots of archived web pages. Continuity spaces are networks of time-consistent persistences; that is, multi-level evolving sets of coherent archived web resources in which researchers can conduct historical analysis at scale.

We illustrate our approach by studying the history of a *dead*[4] constellation of networking web sites: the *firsttuesday* movement (see 3) that acted in the interest of the growth of the '*new economy*' [4] in the early 2000's. Our methodological contribution has practical value for web archives explorers (historians, sociologists, anthropologists, etc.) that aim at conducting large scale analysis of the past web. That is, any diachronic study that goes beyond purely qualitative strategies (see 1.3).

We eventually plan to use web archives for a more systemic purpose. In line with *Complex Systems* approaches, we think that our framework could be a first stone in the study of the *morphogenesis of the web* (see 5). We consider the Web as a complex and dynamic system shaped by socio-technical mechanisms explaining the emergence of online collective forms. We thus aim at using web archives as raw materials to reconstruct the evolutionary structure of the Web at various level of complexity from its origins to the present day.

1.2 Research questions

In order to be able to give a concrete example of application in section 3, we restrict the outcomes of our study to the fields of web history, web historiography and historical research based on web archival materials. Within this application scope, we formulate two broad research questions:

1. How can we reconstruct the evolution of a web site / group of web sites from their full set of archived traces?

2. How can we reconstruct the evolution of an historical event whose traces can be found within archived resources?

These questions complement one another as they investigate web archives from the angle of both archives containers (1.) and archived content (2.). These questions will be enriched with technical observations in section 1.5. Our investigations will also be restricted to the scope of groups of less than 50 archived web sites that no longer exist on the living web[5]. We discuss in section 4 the scalability limits of our method and we question the possible benefits of combining dead, abandoned and living web sites.

[3]Launched in 1996, the *Internet Archive* library keeps the traces of over 747 billion of web pages, see `https://archive.org/`.

[4]By using *dead*, we refer to web sites that no longer exist in the living web according to the terminology introduced by [3].

[5]That is, dead web sites only

1.3 The state of the art of web archiving

Our state of the art will be divided into two stages: we will first conduct a wide review of the research domain of web archiving to clarify our own epistemological choices; we will then focus on key methods and existing results that will be integrated in our own approach (see 1.6).

Mapping the research domain of web archiving. Web archiving as a research domain is almost as old as the web itself: B. Kahle launched the *Internet Archive* initiative in 1996 [5], only 6 years after the invention of the web by T. Berners-Lee and R. Caillaut in 1989-1990 [6]. The first academic papers appeared in 1997 but the domain fully emerged around 2003 after the publication of the *Charter on the Preservation of Digital Heritage* by the Unesco [1]. The core of web archiving literature was mostly published between 2005 and 2015, it touches various domains like Computer Sciences, Digital Humanities, Knowledge Management, etc.

In order to synthesize the scientific landscape of web archiving (sub-fields, trends, research communities, etc.), we have analyzed the vocabulary used in a set of 738 related papers. We have then reconstructed a global semantic map from these extracted terms and expressions by using the free text mining software *GarganText* [7]. Please refer to S1 Appendix for a detailed review of the GarganText method. The resulting map Figure 1 reveals communities of terms frequently used together (the numbered dense area); that is the sub-fields of research that make up the whole domain of web archiving. These sub-fields are numbered by lifespan order and can be labeled as follows:

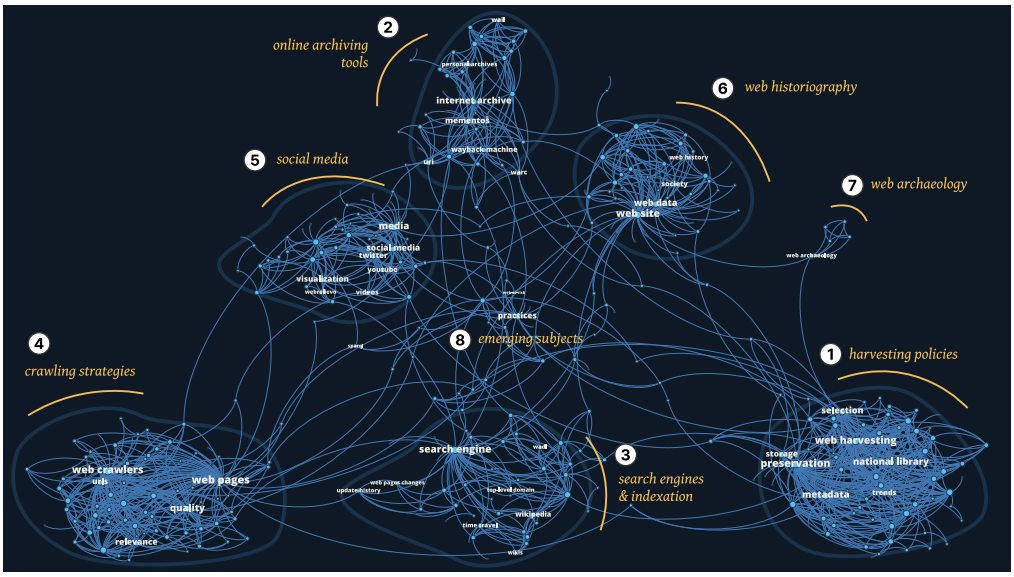

Fig 1. Semantic map of the research domain of web archiving. Created with GarganText from 738 papers metadata (title, date and abstract). Nodes are research topics. The size of a node matches its degree and the size of its label matches its PageRank. Main fields of research are colored and numbered by lifespan order [1..8].

1. Harvesting policies and selection criteria from the point of view of curators, digital preservation and archives accessibility in libraries (1997 - 2010's);
2. Development and uses of online archiving tools, the *Internet Archive*, the *Wayback Machine* (1997 - 2010) and the *Mementos* suite (2010 - 2020);

3. Search engines and indexation of archived resources (2006 - ...) regarding the ephemerality of web pages and their changes frequency (2010's);

4. Improving crawlers and crawling strategies, discussing the induced notions of *coherence* and *quality* of web archives corpora (2009 - ...);

5. Social media and online communities analysis (2009 - ...). This sub-field relies on Sociology and Anthropology but also on Graph Theory;

6. Maturity of web archives analysis with the practical and theoretical emergence of web History and web Historiography (2009 - ...);

7. web Archaeology can be seen as a satellite of web History but more focused on the history of web technologies (2009 - ...);

8. Interdisciplinary and emerging subjects such as gender, languages, online practices, security or ethics (2015's - ...)

After reviewing the Figure 1, we reinforce our intuition that there is a scientific need for large scale explorations of web archives. There are very convincing qualitative or micro-to-mezzo studies in field #6 [8–10] but too few tried to dive into the whole richness of archived corpora. Our own contribution thus aims for weaving an epistemic link between the technical outcomes of fields {#3, #4} and the historical analysis approaches developed in the field #6. We want to investigate the inner structure of web archives corpora – from a data analysis perspective – to address the question of their temporal continuity and discontinuity. By doing so, we will be able to define a *multi-level* exploration framework usable by historians and social scientists. To that end, we will repurpose visualization works from field #5, extend pioneer studies introduced in field #8 and use archived materials provided by the online tools of field #2.

1.4 Towards multiple levels of complexity

Thanks to the Figure 1, we notice that each sub-field of research focuses on a specific *level* of archived resources and doesn't seem to come out of it : *national web* for field #1, *platforms and networks* of web sites for field #5, *web sites* for field #6, *web pages* for field #4 along with *fragmented approaches* shared with field #8. In line with Complex Systems principles, recent works [11, 12] have yet shown that *multi-level* approaches could enrich and refine the outcomes of Digital Humanities research. By choosing a given level of complexity (a web page, a web site, a groups of web sites, etc.), we determine the intrinsic nature of the temporal processes and properties we will be able to analyze. By zooming in and out from one level to another within the same framework, we could be able to understand the whole and complex network of dynamic processes that make up the historical subject we try to reconstruct. Our contribution is thus wider than proposing a quantitative study of web archives: we seek to conduct *multi-level* explorations.

1.5 The discontinuity problem

The paper [3] has proven the following statement: '*web archives are not direct traces of the web, they are traces of crawlers*'. This fact explains most of the issues encountered while exploring web archives at scale from inconsistency to redundancy and temporal blindness. In other words: discontinuity issues are induced by the technical nature of web archives. To clarify the discontinuity problem, we define a minimalist web archiving model inspired by [13] and illustrated with the Figure 2. There, a web page p evolves along a continuous timeline. This page occasionally mutates over time by adding,

modifying or removing content. It results in a series of changes $\mathcal{T}^* = \{t_i(p)\}_{1 \leq i \leq l}$. We here make the assumption that we don't know the nature of those changes. We then call $\mathcal{C}^* = \{C_j\}_{1 \leq j \leq m}$ an ordered series of crawls periodically scheduled to harvest p and catch most of its mutations. Each crawl C_j goes with a lifespan $\Delta_{C_j} = [t_j, t'_j]$ with $t_j < t'_j$ and produces exactly one snapshot of p timestamped such as:

$$\forall p, \forall j,\ C_j : p \mapsto c_j(p) \text{ with } t_{c_j(p)} = t'_j$$

Here $c_j(p)$ is a snapshot taken by C_j and associated to a downloading date $t_{c_j(p)}$.

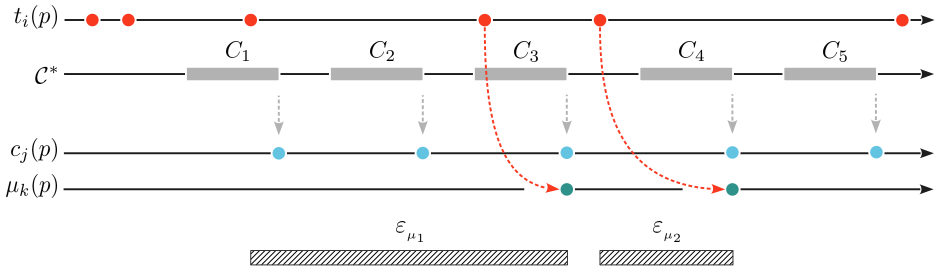

Fig 2. A minimalist web archiving model where red dots are page's mutations, blue dots are archived snapshots resulting from a crawl and green dots are known page's mutations deduced from each crawl.

The continuous temporality of p is discretized by each crawl as p is reduced to a finite collection of snapshots. This harvesting process binds the resulting archive to an arbitrary downloading date disconnected from the original mutations of p. From that point, our knowledge of the page evolution will strictly be determined through the mutations caught by the crawls. We call $\{\mu_1(p), ...\mu_k(p)\}$ the set of known changes which can be inferred based on crawls[6]. The dating of each known change goes with an imprecision $\varepsilon_{\mu_k} = \mu_k(p) - t_i(p)$.

Redundancies in web archives are induced by over-harvesting unchanged pages. Temporal blindness comes from not having captured one or more page's mutations. Inconsistency occurs when we put in relation two archived pages p_1, p_2 collected by two different crawls without knowing if one of them has evolved over the time span. Checking such an inconsistency is only feasible by hand for corpora of less than one hundred pages. Beyond this limit, the discontinuities are such that no existing method can overcome them. The discontinuity problem of web archives thus results from the combinations of redundancies, temporal blindness and inconsistency at the scale of entire web sites, of communities of blogs, of larger sub-spaces of the web, etc. Furthermore, this problem is highly increased by the timestamping imprecision ε_{μ_k}.

To sum up, our main challenge will be to track remaining pieces of continuity within existing corpora of web archives and to reconstruct explorable topographic spaces (see 2). Such continuity spaces will depict the evolution of archived resources and highlight sub-area of continuity where multi-level analysis might eventually be conducted. In order to fulfil this goal, existing works will be re-used or extended.

1.6 Related work

Web archives coherence and quality. As we will deal with more than a single archived page, we now need to extend the previous model with the notions of *coherence*

[6] An archived known change is detected if the content of p has evolved between two consecutive crawls.

and *quality*. Each of these notions have already been described by a distinct community of scholars: *coherence* has been investigated in relation to search issues (field #3) and *quality* has been defined regarding harvesting purposes (field #4). The notion of coherence induces the presence of an observer that navigates through the archives by means of a search engine. As the observer moves from one archived page to another, the notion of coherence can be understood as a question of temporal consistency between two visited archives. For its part, the notion of quality is a crawler-side metric: it measures the completeness of a crawl regarding the current structure of a web site in the living web. The work of M. Spaniol [13] made a connection between coherence and quality by defining the extended notions of *observable coherence* and *observable quality*. In what follows, we will reuse this two notions as quality scores to measure the temporal and structural consistency of our own reconstructions.

Web fragments. *Elementary approaches* are sub-research fields of harvesting issues (field #4). They aim for improving both *focused-crawl strategies* and the resulting quality of the archived corpora by collecting more meaningful historical content for future research. Various models have been successively introduced by the community: visual elements extraction [14], blocks [15], segments [16] and main content targeting [17]. The most recent model – the *web fragment* framework [3] – has the particularity to take the point of view of an observer into account and to be usable in an exploratory context. The web fragment is a disaggregated level of analysis where archived pages are split into coherent fragments and timestamped by *edition date*. This interdisciplinary work creates a direct connection between field #4 and field #6 by implementing the theoretical notion of *analytical web strata* [18]. In addition, the web fragment framework extends the notions of *observable coherence* and *observable quality* by defining the notions of *disaggregated observable coherence and quality* [3]. This framework is a key contribution for our own work as we aim for reconstructing the evolution of archived resources at different levels of complexity.

Web Archives evolution and visualization. Due to discontinuity issues, using web archives corpora to visualize the evolution of web resources at scale was not deeply explored by scholars. The most convincing diachronic analysis were conducted from a qualitative historical or historiographical perspective such as: the study of the 90's '*web of pros*' [8], the observation of diasporic online communities [10] or the analysis of dying web sites [19]. Tools used to fulfill these studies were search engines parameterized for browsing single archived web sites [20] or pages [21]. Quantitative approaches mostly relied on *temporal series of hyperlinks graphs* in the wake of the pioneer works of M. Toyoda and M. Kitsuregawa [22]. These approaches gave birth to practical tools like *webverse*[7] or *Linkgate*[8]. The *hyperlink analysis* method was also used to study individual web sites [23], large communities [24] or national sub-spaces of the web [25–28]. Some studies smartly bypass the complexity of studying archives evolution by using *visual data analysis* [9, 29], *text analysis* [30] or *crawler's log analysis* [31]. Despite having played a key part in the history of web archives analysis, these past works cannot be generalized (or hardly) to other research questions as they remain connected to a specific topic, a given level of complexity or a sub-element of archived resources (image, title, links, etc.). In this paper, we aim for building a agnostic and multi-level methodology of exploration that could be applied to and from various contexts. For this reason, we are particularly interested by the work of B. Fry: the *Anemone* project [32]. As far as we know, this visualization is the first and only attempt to reconstruct the dynamic evolution of a single archived web site in a generic

[7] See `http://webverse.org/`.
[8] See `https://netpreserve.org/projects/linkgate`.

perspective. The *Anemone* results in an explorable graphical space where one can witness the emergence of entire sub-parts of an archived web site over time. We will use this work as inspiration for our own contribution.

Phylomemies. Phylomemies are inheritance networks of textual elements [33, 34] designed to reconstruct the evolution of semantic landscapes from raw texts. The *phylomemy reconstruction process* thus combines advanced text-mining methods, scientometrics, visualizations, and methods for the reconstruction of evolving complex networks to reconstruct the latent semantic structures of an unstructured but timestamped set of textual documents. In what follows, we will reconstruct phylomemies from web archival materials and in order to study the history of online debates and the evolution of topics within dead web sites.

2 Materials and methods

Our article intends to resolve the discontinuity problem of web archives and then to put our solution into practice by analysing a real set of archived resources. Of course, we want our case study and experimental material to be of interest for historians. We have thus decided to explore the archives of the dead web site *firsttuesday.com* due to its historical importance for digital capitalism and because of its complex structure as a web portal that gathered online and offline communities of entrepreneurs and investors in the late 90s.

2.1 Experimental material: the archives of *firsttuesday.com*

The year 1995 was a turning point in the history of the Web: the Initial Public Opening of *Netscape* in August 1995 marked the beginning of the 2000's *dot-com bubble*. The expansion of this *New Economy* came out of the idea that Internet and the Web could create new type of business markets and achieve unprecedented returns on investment [4]. The Web quickly became the source of a financial euphoria and a gold rush. By that time, venture capital was massively available and valuations in startups related to Information and Communication Technologies (ICT) followed an exponential growth. Starting an online business required almost no capital base and consequently startups popped up all over the USA and Canada before reaching Europe [35]. Stock options over-increased the capitalization of young companies and venture capitalists lavishly spent their investments in search of short term profits. Between 1995 and 2000, the Web was no more the territory of nerds and inventors, it was the playground of entrepreneurs and investors: one could easily make millions of dollars on the basis of an attractive business plan. The dot-com bubble crashed in March 2000 and dragged the global valuation of tech-markets down with it. It took almost ten years – and the success of *Facebook* – for confidence in the digital markets to return.

Created in 1998 in Great-Britain, the *First Tuesday* events played a key role during this period of euphoria. Every First Tuesday evening of each month, big meetings gathered hundreds of investors and entrepreneurs together in prestigious places (luxurious hotels, headquarters of companies, ministries, etc.). Keynote lectures given by renowned startup founders were announced online on *firsttuesday.com*. But beyond those speeches, coming to a First Tuesday event was the occasion for entrepreneurs (marked with a yellow badge) to meet investors (marked with a green badge), to put on their business plan and eventually to raise funds. For a few hours, these events turned themselves into giant ephemeral social networks. In the early 2000s, the concept of the First Tuesday spread throughout North America and Europe in the form of regional and

local chapters. The First Tuesday events peaked in 2001 before slowly decreasing and becoming confidential as born digital networking platforms like *LinkedIn* rose.

Despite their historical importance, the First Tuesday events have not yet been studied by scholars and, as far as we know, there are no records left of these meetings. However, we think that some valuable information can be found in the archives of the web site *firsttuesday.com*. Indeed, it acted as a hub for the whole community with ads for upcoming events (date, place and description), reports of past meetings, a discussion forum and a list of local chapters[9].The archives of *firsttuesday.com* might be used to understand the staging of the narratives of the 2000s web economy (heroization of business leaders, entrepreneurial success stories, etc.) and study, by extension, the genesis of the myths of today's digital capitalism. The symbolic world of startups has been driven by the invention of a common language ("founders", "stock-options", "success", etc.) and by the faith in shared dreams and utopia like the figure of the *self made man* [36].

As an experimental protocol, we have asked historians for research questions related to the historical context and content of the *firsttuesday.com* web site. These questions will be addressed one by one in section 3 :

- How did the structure of *firsttuesday.com* evolve throughout time? Have any new sections been created? Can we follow the evolution of its graphic design? Can we qualify the nature of the interactions between users inside the platform?
- Can we reconstruct the development of the First Tuesday constellation? What were the differences between the regional chapters and the main web site? Were there any local particularities?
- Can we map the dissemination of the First Tuesday events over time and space?
- Were there temporally, geographically or thematically structured networks of sponsors and speakers?
- Can we study the structure and dynamics of the vocabulary used during the meetings? Can we reconstruct the semantic landscape of the New Economy?

Data sets. The archives of *firsttuesday.com* have been extracted from the *Internet Archive* database by using the Wayback CDX Server API. We have first harvested a collection of 8280 snapshots before reducing them to set of 3670 de-duplicated captures that represent 1507 unique archived web pages. This data set is called D_1 and covers a period of 12 years from the first capture of *firsttuesday.com* in 1999 to its death in 2010. By quickly analysing the HTML content of D_1, we find that at least 203 pages[10] haven't been preserved by *Internet Archive*. The loss of these web resources especially concerns the online discussion forum of *firsttuesday.com*. We next use D_1 to reconstruct the citation neighbourhood of *firsttuesday.com* by following its outgoing hyperlinks. We divide this neighbourhood in two distinct data sets: the *First Tuesday constellation* called D_2 that gathers a list of 101 web sites whose domain names contain the term "firsttuesday" apart from the main portal *firsttuesday.com*; and D_3 a corpus of 1034 web sites whose domain names do not contain "firsttuesday". The data set D_2 gathers the regional chapters of *firsttuesday.com* (standalone sites, *yahoo groups* and *e-groups*) while D_3 mostly assembles the technological and financial sponsors of the events[11].

[9]Those local places were either standalone sites – quasi-siblings of the original *firsttuesday.com* – or simple discussion groups hosted by larger platforms (*yahoo groups*, *e-groups*, etc.)

[10]An un-archived page is a web page cited in D_1 and whose url starts with the prefix *firsttuesday.com* but that is not part of the *Internet Archive* database.

[11]The corpus D_1, D_2 and D_3 can be downloaded at `https://doi.org/10.7910/DVN/EAFOHY`

2.2 Software

For requirements of repeatability and accessibility, we choose to use free and open-source technologies. We aim not only to build software that can be deployed on large clusters of servers but also to let non-computer specialists use these technologies from their own personal laptops for small-to-medium scale analysis. We have thus developed a set of generic Python scripts – the *Déndron*[12] suite – for harvesting web archives (2.1), detecting *persistences* (2.3.1) and building *continuity spaces* (2.3.2). Déndron also goes with a web interface designed to visualize and explore *web cernes* (2.3.5). See S2 Appendix for implementation details. In addition, we use the free text mining software *GarganText* [37] for analysing the textual content of the archives and for reconstructing their temporal dynamics (3.5).

2.3 Method of reconstruction of continuity spaces

The empirical cartography of parts of the living Web, popularized in the 2010's by software like *Gephi* [38] and research projects like the *e-diasporas Atlas* [39], has proven that researchers could make use of graph visualization to study the structure of the Web and analyse the complexity of online social relationships at a given moment. By following a similar approach from web archives, we could add a temporal dimension to our investigations and thus increase our capacity to understand the dynamics of digital social systems. Unfortunately, we have shown in subsection 1.5 that discontinuity issues induced by crawlers prevent us from conducting such analysis at large scale. In what follows, we will introduce a method to rearrange sets of archived pages in the form of *persistences*. These persistences will then be used to detect, reconstruct and visualize *continuity spaces* within corpora of web archives and eventually help us to address historical research questions as listed in subsection 2.1.

2.3.1 Persistences

We aim at re-introducing a form of duration in the discretized temporality of web archives corpora. According to the philosopher H. Bergson, the concept of duration can be defined as "*a time perceived and lived*" that "*coexists in consciousness with a series of successive states*" [40]. In other words, the duration is a phenomenological construction that induces the presence of a subjective observer. The analogy of this definition with the model depicted by the Figure 2 is clarifying. Indeed, while browsing web archives, we experience what remains of the past Web through the perception of crawlers in the form of finite collections of snapshots. We thus think that the key to eventually reconstruct *continuity spaces* is to define such a notion of duration applied to web archives. In particular, we need to focus on subsets of durations called *persistences*.

For a given archived web page p, we define a persistence τ_n as a temporal sequence of unchanged snapshots $c_j(p)$ such as:

$$\tau_n = (c_j(p))_{1 \leq j \leq m} \text{ with } c_j^{\prec}(p) = c_j^{\succ}(p) \text{ and } t_{c_j^{\prec}(p)} < t_{c_j^{\succ}(p)}$$

In extreme cases, if the crawler archived p only once, τ_n will take the form of a single snapshot; that is, a persistence without duration called *impermanence*. If the page didn't change during the entire crawling campaign $\mathcal{C}^*$, it will results in a single long-duration persistence. On the contrary, if p changed between every crawl there will be exactly m impermanences. So the nature of τ_n depends on the quality of the

[12] *Déndron* is downloadable at `https://gitlab.iscpif.fr/qlobbe/dendron`. Licence : AGPL + CECILL v3

crawling campaign as illustrated with Figure 3. In what follows, we will focus on persistences that are not impermanences[13].

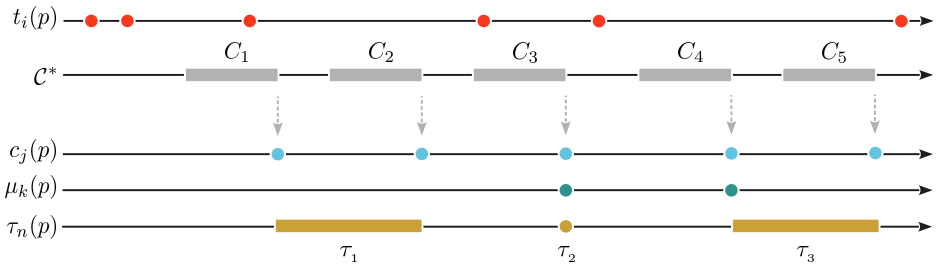

Fig 3. Persistences (brown lines) and impermanences (brown dots).

Multi-level persistences. Persistences can be defined at various levels of complexity by aggregating sets of archived pages based on structural rules (domain names, hyperlinks, etc.) or qualitative choices (common themes, historical periods, social groups, etc.). By doing so, we meet the multi-level requirements formulated in section 1.3 and move from the pages' granularity to the larger level of an ad hoc aggregator called $\mathcal{P}$.

The global aggregation function Γ can be decomposed into three operators that correspond to three main steps $\Gamma = {}^3\gamma \circ {}^2\gamma \circ {}^1\gamma$. The first operator ${}^1\gamma$ is a selection function that acts at the page level. It shades the minimal model of Figure 2 by allowing the researcher to choose whether a known change $\mu_k(p)$ has to be taken into account or not. It qualifies the nature of those changes regarding a given research question as discussed in [41,42]. The second operator ${}^2\gamma$ constructs persistences from the results of ${}^1\gamma$ inside each page as defined by Figure 3. The last operator ${}^3\gamma$ is a merger function that aggregates sets of persistences $\{\tau_n(p)\}_{p\in\mathcal{P}}$ into a single series $\tau^*(\mathcal{P})$. This merger is based on a set algebra between persistences and/or impermanences. In fact, the concrete configuration of Γ depends on each research question. In 3.2, we give an example of such an aggregation at a web site level.

Continuity links. At the page level, if their durations overlap, a *continuity link* can be drawn between two persistences belonging to two different pages p_1 and p_2 such as:

$$L_x(1,2) = \{\tau_n(p_1), \tau_{n'}(p_2)\} \text{ with } [t_1, t_1'] \cap [t_2, t_2'] \neq \emptyset$$

There, $[t_1, t_1']$ is the duration of $\tau_n(p_1)$ and $[t_2, t_2']$ is the duration of $\tau_{n'}(p_2)$. The continuity link $L_x(1,2)$ is undirected and there might be several links between p_1 and p_2 as explained with Figure 4. At a higher level of complexity, with two aggregators $\mathcal{P}_1$ and $\mathcal{P}_2$, continuity links will take the form of $L^*_{x'}(1,2) = \{\tau^*_n(\mathcal{P}_1), \tau^*_{n'}(\mathcal{P}_2)\}$.

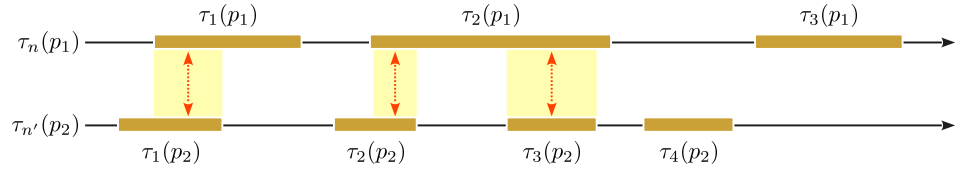

Fig 4. Three continuity links (red arrow) are drawn between the persistences of pages p_1 and p_2 as their durations overlap (yellow area).

[13] We however note that it might be interesting to focus on sequences of impermanences to study moments of high activity in an archived web page.

2.3.2 Graph of continuity

Persistences have so far been defined as unidimensional objects. We now combine them to create explorable continuity spaces by using a geometric approach. A *graph of continuity* is thus a network of persistences and continuity links defined as $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ with $\mathcal{N} = \{\tau_n(p)\}_{p \in P^*}$ and $\mathcal{L} = \{L_x(p, p')\}_{p,p' \in P^*}$ where P^* represents a coherent subset of archived pages, for instance all the snapshots that belonged to the same domain name.

But such a graph could quickly become very dense as we might attempt to weave all the possible continuity links between elements of P^* with a maximal computational complexity of $O(n^3)$. We however don't need to compute all those links. Indeed, we want $\mathcal{G}$ to respect the historical browsing experience of P^* in the same way crawlers and web users did years ago. That's the reason why we choose to add series of constraint operators $\{^y\varphi\}$ design to spatially or structurally control the making of the graph such as $\mathcal{G} = (^y\varphi \circ .. \circ {}^2\varphi \circ {}^1\varphi) \circ (\mathcal{N}, \mathcal{L})$. For instance, in section 3.1, we only draw continuity links between archived pages connected by hyperlinks or between sibling pages belonging to the same stage in the tree like URI structure of *firsttuesday.com*.

2.3.3 Continuity spaces

Graphs of continuity can be studied through the lens of graph theory to extract temporally and structurally coherent sub-spaces such as cliques, connected components and cycles:

- cliques are subsets of nodes such that every two distinct nodes are adjacent. In our model, a clique embodies a *space of complete continuity*;
- connected components are connected sub-graphs where nodes are linked together through (at least) one path. Within graphs of continuity, connected components can bee understood as *spaces of partial continuity*;
- cycles are non-empty trials in which first and last nodes are equals. Provided that crawls are of good quality, cycles will follow and interlock over the course of time. At the level of a given archived web site, breaks between cycles will point out major changes in the history of the site's structure: arrival of a new CMS, global redesign of a graphic chart, etc.

Continuity spaces are new scientific objects in the research domain of web archiving. They are to become the subject of studies dedicated to their topographic properties (density, size or vicinity ; see 3.1). They will also serve as reliable starting points for data driven historical research, they will be stable fields from where archived traces of past phenomenon will be extracted (see 3.5).

2.3.4 Temporal and structural quality

The corollary of the existence of continuity spaces is that the temporal and structural quality of any historical analysis of web archives will now be measurable. Indeed, prior to any analysis, relevant set of archived snapshots are selected and extracted from the main corpora. This selection obviously shapes the quality of the upcoming results. And we are entitled to ask ourselves: what would be the scientific quality of an analysis based on totally incoherent elements?

This question can now be addressed through continuity spaces by projecting the input data of any analysis onto these area. Are the input snapshots all part of the same continuity space? Are they on the crossing of different cliques? Are they strictly composed of unconnected spaces? etc.

To clarify this idea, we define a temporal and structural quality function inspired by the "*crawl quality*" function of M. Spaniol et al. in [13]. We first refer to $\mathcal{A}(\mathcal{P}^*)$ as a given historical analysis. We call $\mathcal{P}^*$ the set of archived pages used as research material by $\mathcal{A}$. We denote by $\mathcal{S}^*$ the complete set of continuity spaces reconstructed from $\mathcal{P}^*$. Any $\mathcal{S}_i \in \mathcal{S}^*$ is a graph $\mathcal{S}_i = (\mathcal{N}_i, \mathcal{L}_i)$ whose nodes $\mathcal{N}_i$ are persistences τ_i^*. In what follows, we restrict the comparison of two continuity spaces to the comparison of their nodes[14]. The most simple and classical measurement that can be used to compare two sets of elements is the Jaccard index. Given two spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ reduced to sets of nodes $\mathcal{N}_1 = \tau_1^*$ and $\mathcal{N}_2 = \tau_2^*$, the Jaccard index is defined by:

$$J(\mathcal{S}_1, \mathcal{S}_2) = \frac{|\mathcal{N}_1 \cap \mathcal{N}_2|}{|\mathcal{N}_1 \cup \mathcal{N}_2|} \text{ such as } J(\mathcal{S}_1, \mathcal{S}_2) \mapsto [0, 1]$$

A broader similarity function can be defined to take into account subspaces:

$$D(\mathcal{S}_1, \mathcal{S}_2) = \begin{cases} 1, \text{ if } \mathcal{S}_1 \subseteq \mathcal{S}_2 \text{ or } \mathcal{S}_2 \subseteq \mathcal{S}_1 \\ J(\mathcal{S}_1, \mathcal{S}_2), \text{ otherwise} \end{cases}$$

Finally, the overall temporal and structural quality of an analysis $\mathcal{A}$ is defined by:

$$\mathcal{Q}(\mathcal{A}) = \frac{\sum_i^n \sum_j^n D(\mathcal{S}_i, \mathcal{S}_j)}{n(n-1)}$$

This quality function can be used to improve the repeatability of exploration methods of web archives and support the development of this field as a scientific domain: two results obtained from a same collection of archives can be compared, a given exploration method can be challenged regarding its quality, the result of an analysis can be improved by merging multiple corpora given the evolution of the quality function, etc. This measure can also be used during the exploration step to help researchers choosing the best level of complexity ; knowing that this function can be applied to various granularity.

2.3.5 Web cernes and dendro-reconstruction

Persistences can be used for visualization purposes as they reflect the structural evolution of archived web sites. In what follow, we aim at projecting a graph of continuity into a graphical space and interact with it. We thus want to come up with a meaningful representation able to translate the local and global dynamics of a given web site (emergence, growth, adaptation, phases, etc.) and reveal the way it organized itself into evolving sub spaces. This new type of visualization will be designed to reduce a web site made of hundred of pages into a single and significant shape comparable to other.

But as explained in 1.3, the question of the visualization of web archives remains an open subject. Previous works have mostly chosen to make the tree-like structure of archived URLs appear in their end visualizations even if those trees are much more the reflection of an internal skeleton than the traces of an external shape. Indeed, the mental image built by a web user while browsing a web site is midway between a hierarchical and an organic representation [32]. In our proposition, URLs trees will be used in the background to guarantee a structural coherence to our visualization and only persistences and continuity spaces will be visible.

In the same way as [32], our inspiration comes from nature and in particular from tree-rings also called *cernes*. Tree-rings are horizontal cross sections cut through the trunk of a tree and occurring in layers. Tree-rings give evidence of yearly trees growth

[14] Future developments will be dedicated to the review of more precise comparison measures [43] able to with all the complete topographic properties of graphs (nodes and links).

and can be used to date them (i.e., the dendrochronology) or to study climate evolution from wood (i.e., the dendroclimatology).

We call our visualization technique *the web cernes*. In the web cernes, the bark of a web site (its external shape) results from the growth of each archived pages along anamorphic timelines. These pages are organised around a central axis by following the *dendro-reconstruction method*. This method is illustrated by the Figure 5 and can be decomposed as a six steps process:

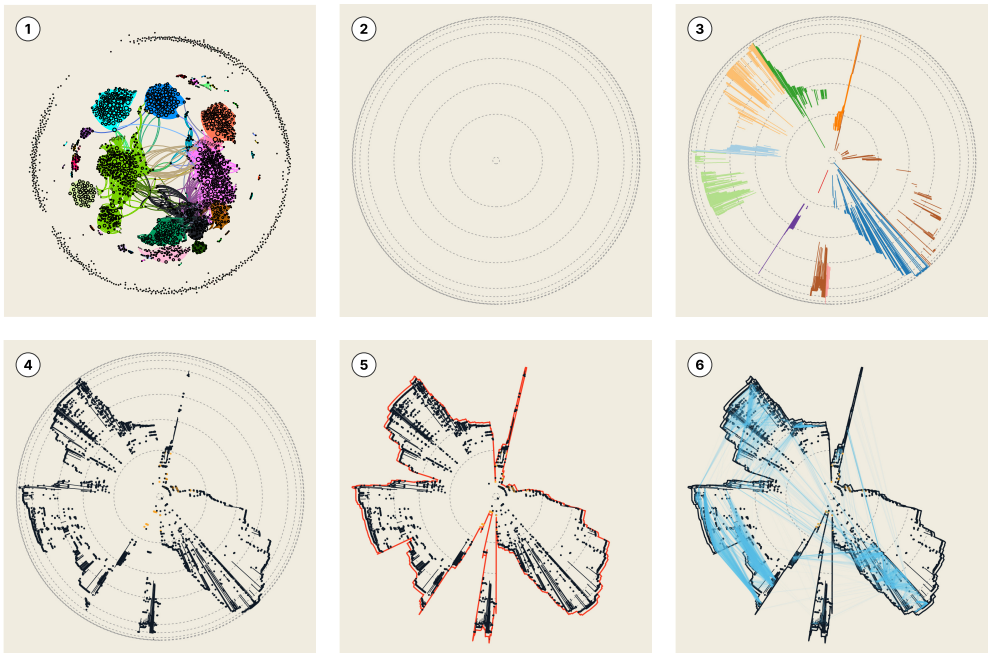

Fig 5. The six steps of the dendro-reconstruction method to produce web cernes

1. We first harvest all the archives of the targeted web site. We aggregate them by URL and extract references to un-archived resources (see 2.1). This primary corpus is called $\mathcal{P}^*$. We next reconstruct persistences and impermanences from $\mathcal{P}^*$ and compute their continuity graph $\mathcal{G}$ along with continuity spaces $\mathcal{S}^*$. We then aggregate the nodes of $\mathcal{G}$ regarding their respective URL and finally use the Louvain method [44] to split this aggregated graph into large partitions $\{g_i\}$ (the coloured dense communities in Figure 5.1).

2. At the same time, we use the web fragment method [3] to ascertain the creation date of every page. We count the number of new pages created per year and draw in Figure 5.2 a series of concentric dotted circles for each of these years. The center of the figure called O thus corresponds to January 1st of the oldest year and the distance between to subsequent years t_1 and t_2 matches the number of pages created throughout t_1.

3. We reconstruct the URLs tree from $\mathcal{P}^*$ and use it as a reference to sort the partitions $\{g_i\}$. Indeed, each of these partitions can be moved closer or further away from each other according to the distance between their respective pages in the URLs tree. We thus end up with a sorted series of $\{g_i\}$ that can be projected into the visualization and distributed around the pole O. The partition that includes the home page of the targeted site is put at the upright top position of O. Following the same procedure, the pages are then sorted within each partition

before being associated to a set of polar coordinates. We finally draw a line for every page (the coloured lines in Figure 5.3).

4. From a graphical perspective, each page spreads along a dedicated line bounded by the page's creation and last archiving dates. Those lines can indeed be seen as anamorphic temporal axis shaped by the graduation of the dotted circles. We next replace each lines by their corresponding page's persistences and impermanences (the black lines and dots in Figure 5.4). Un-archived resources are depicted by orange dots.

5. We draw an external curve (the red line in Figure 5.5) by joining the end of each individual page's line. This curve contains all the graphical elements and thus finalizes the shape of the web cernes.

6. We project the continuity spaces (the blue lines in Figure 5.6) to highlight areas where large scale analysis can be conducted without inconsistency.

3 Results

Through the notion of persistence, we develop an analytical framework that turns raw archival materials into explorable visualizations. We apply this framework to the historical study of the archives of *firsttuesday.com*. We thus address each research questions formulated in 2.1. We investigate the online/offline federal strategy of the *firsttuesday* ecosystem. We analyse its rise and death in relation to the crash of the dot-com bubble in the early 2000's. We show how the *firsttuesday* events acted as melting pots for the development of digital capitalism's narratives. All our analysis are based on the exploration of the corpus D_1.

3.1 The structural evolution of the web site *firsttuesday.com*

The first question asked by historians in 2.1 was: *How has the structure of firsttuesday.com evolved throughout time?* To answer this question, we first calculate the persistences from D_1 before reconstructing the web cernes. The Figure 6 thus represents the evolving shape of *firsttuesday.com* within which twelve consistent spaces are numbered and coloured.

The platform was mainly active between 2000 and 2003 with a growth peak during the winter 2000 - 2001 ($28,4\%$ of its 1657 archived pages were created between January and December 2000). Its activity quickly decreased after 2004. We notice the loss of the discussion forum (space #11 in Figure 6). Created in April 2000, this section was unfortunately never archived. On the whole, the structure of *firsttuesday.com* was stable over time as we found no major discontinuity in the *home* and *menu* pages area (space #1). In other words: the main structuring pages had just to undergo minor evolutions. The only noticeable change affected the editorial content of space #6 that moved from main pages to a dedicated section called *content* in 2001. The space #2 bears witness of an ephemeral attempt to promote alternative events called *Wireless Wednesday*. The descriptions of those *mobile* related events moved to a dedicated web site named *wirelesswednesday.com* in the late 2001 before disappearing in 2003. The archived content published on *firsttuesday.com* can be divided into two categories: the standalone and static editorial content created by the web site team (spaces #1, #6, #10) and the content that gathered information from the entire *firsttuesday* ecosystem (ie, the local *chapters* and *events*). The second category gave form to the most important part of the site. Local and regional *firsttuesday* chapters were described within space #3. There, one could find links to the standalone web sites of each chapters along with

practical informations (scheduling of local events, organization chart, etc.). An interactive map was set up in 2003 (space #4) in order to visualize and synthesize all the existing worldwide chapters. New chapters were somehow crowdsourced by means of registration forms (space #5). Thus, *firsttuesday.com* was more a platform for the federation and organisation of an decentralized community than a focal point from which the movement was governed. This unifying will was embodied by way of a global agenda of *firsttuesday* events gathered inside spaces #7, #8, #9 where meaningful informations were shared: date, place, participants, speakers, etc.

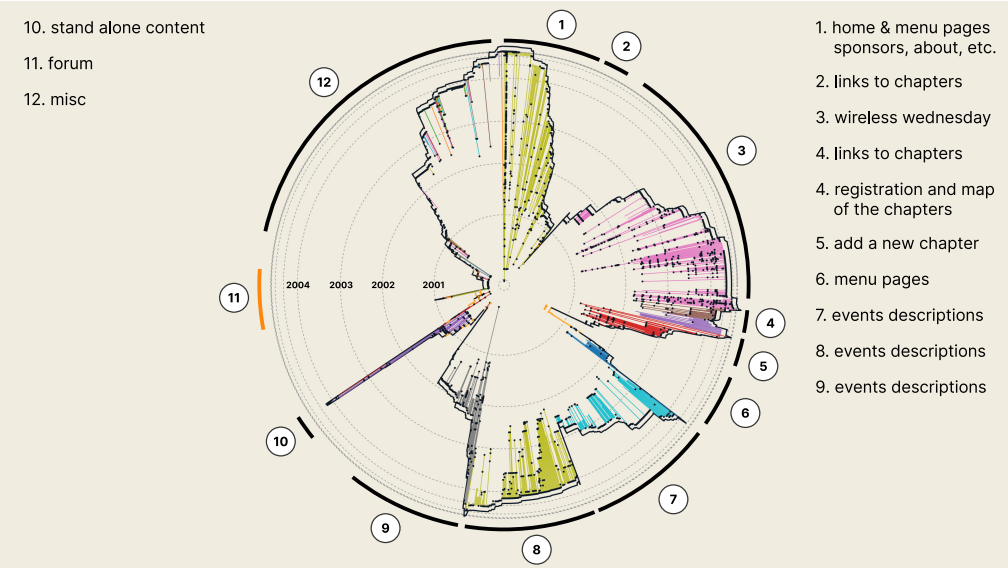

Fig 6. Web cernes reconstructed from the archives of *firsttuesday.com*. The colored numbered areas feature spaces of continuity. Explore the interactive version.

The union of spaces #7, #8, #9 forms an interesting corpus that can be used to analyse the *firsttuesday* events from an historical perspective. We call this corpus D_4. Its temporal and structural quality is of 0.215 which is significantly better than the rest of the site ($\mathcal{Q}_{D_1 \setminus D_4} = 0.057$). The figure 11 in Appendix thus shows the dense placement of continuity cliques over D_1.

3.2 Rise and death of the *firsttuesday* constellation

In a second phase, historians asked us to extend our analyses and study the evolution of the whole *firsttuesday* constellation; that is, all the archived web sites related to the *firsttuesday* initiative that were gravitating around the main platform.

To do this, we use the corpus D_2 (see 2.1) that contains a list of all the out-citation links of *firsttuesday.com* mentioning "firsttuesday" in their own domain names. We restrict this list to the standalone web sites only and remove the *yahoo groups* and *e-groups*. We end up with a constellation of 37 web sites[15]. For every site, we harvest the archived pages, we calculate the persistences and we reconstruct the web cernes. We next collect all the in/out-citation links between the members of the constellation to shape a network of web cernes. We spatialize[16] this network with Gephi. The Figure 7 reveals the evolution of the *firsttuesday* constellation between 1999 and 2010. This representation can be studied at a global level or at a site level by zooming in within the web cernes.

[15] 34 sites have been archived by Internet Archive, 3 have not.
[16] We here use the *force atlas* algorithm [38].

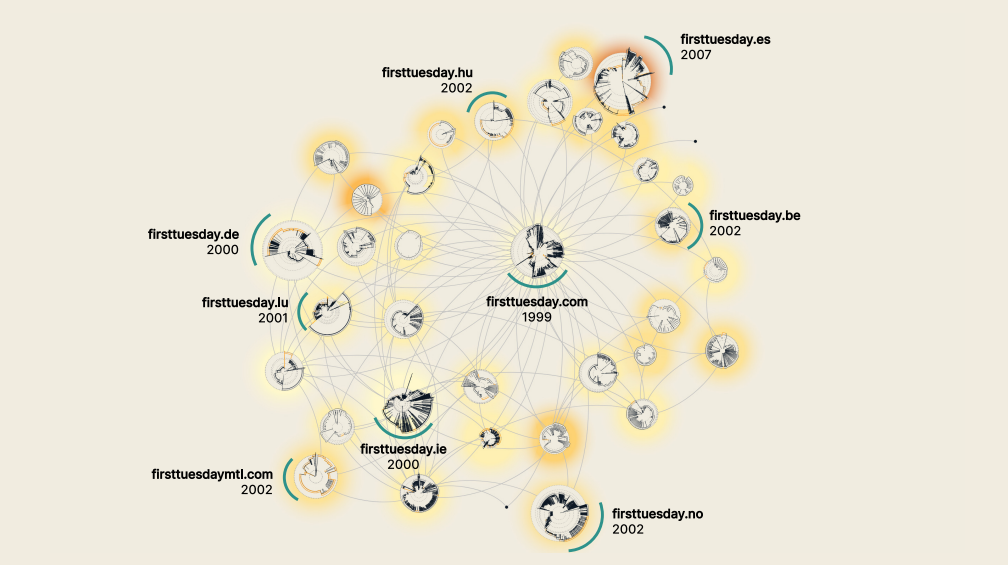

Fig 7. Online network of the *firsttuesday* constellation. 13237 archived pages are rearranged into web cernes that feature the evolution of 37 web sites. Grey paths are in/out-citations links. The network is spatialized with Gephi. White to orange shades map the creation date of each site. Black dots are unarchived sites. Explore the interactive version.

Created in 1999, the main platform *firsttuesday.com* is the oldest site of the constellation and obviously occupies a central position in the network. Its neighbours can be divided in two categories in light of their connectivity.

Poorly connected sites with degree < 3 are spatialized in the upper-right part of the graph[17]. These sites are only connected to the main platform and are among the youngest ones of all the constellation. They bear witness of the attempt of the movement to locally survive or revive itself despite its worldwide decline in 2002.

The lower-left part of the network is much more connected and seems to maintain a form of autonomy from the main platform. They connect to each other without necessarily going through *firsttuesday.com*. This fact reinforces the hypothesis that *firsttuesday.com* was first and foremost an aggregator, a showcase site for the community and not a top down organiser. At a global level, this constellation does not differ from other early 2000s online networks [45]. They share common characteristics like: being a sparse graph (its density is of 0.15), having a distribution of sites' degree that follows a power law (see the figure 12 in Appendix) and being structure around a giant connected component. From a graphical, structural and editorial point of view, each site of the constellation can be seen as a sibling of the main platform. The *.co.uk*, *.sk*, *.de* and *.lu* sites are thus quasi clones of *firsttuesday.com*. None of them tried to differentiate itself. On the contrary, this common design seems to have been a mark of belonging to the community. We nevertheless have found some evolutionary particularities:

- in the wake of the Web 2.0, the Hungarian chapter set up in 2004 a blog section with the possibility to comment news and events ;
- the Belgian chapter staged the creation of a an online business by way of a FAQ section with some storytelling ;

[17]The sites with extensions *north.com*,*.es*,*.br*,*.be*,*scotland.es*,*frankfurt.com*,*cincinnati.com* and *krakow.pl* are considered as poorly connected.

- the Norwegian and Czech chapters briefly revived the *firsttuesday* concept by massively re-organizing local events in 2003 ;
- the Italian chapter led a double life. Before 2002, its activity was similar to the rest of the constellation (with events and editorial content). After 2009, the site tried to reinvent itself by becoming a news platform and a *LinkedIn* like dating site between startup founders and investors ;
- the Luxembourgish chapter long survived through its discussion forum ;
- after a long period of inactivity the English and German chapters lost their domain names in 2006 to be replaced by miscellaneous online adds.

Except for the Spanish chapter (created in 2007), the activity of the constellation decreased in 2003 and the network almost died after 2004. Afterwards, the survivors tried to follow the technical evolution of the web (blogs, professional dating app and social network). But the strength of the *firsttuesday* initiative came from the vivacity of its offline community and from its capacity to act as a worldwide network. After the collapse of the community, the remaining isolated chapters failed to locally revive the *firsttuesday* concept.

3.3 Tracking the geographical evolution of events through time

The organization of events between startup founders and investors was a the heart of the *firsttuesday* initiative especially between 2000 and 2002. As the corpus D_4 turns out to be of good quality compared to the rest of the archives and as it aggregates most of the events published in the rest of the constellation, we now dive into the archived pages of D_4 in order to address the third question asked by historians: *Can we map the dissemination of the firsttuesday events over time and space?*

We here use the *web fragment* framework (see 1.6) to extract from D_4 and date with precision as much events descriptions as possible. An event description may include in its raw text part or the totality of the following informations: a date, a place, a list of invited speakers, a list of sponsors and a description of the topics to be discussed. We thus extract and de-duplicate 593 event descriptions from D_4. Among them[18], 554 can be precisely geolocated at a city scale from 2000 to 2005. Events that happened before 2000 seem to have not been archived or even published online. The 8 represents the evolution of the geographical distribution of these events.

The temporal distribution of these events matches the structural dynamics of *firsttuesday.com* which confirms that the online activity of the site was mostly the reflection of the offline development of the community. On the whole, the events took place in the Western world starting with the United States and then Europe. In the USA, the events gathered in the West Coast and global West around the technology and financial centres of the early 2000s (Chicago, Washington DC, Cincinnati, Miami, etc.). In Europe, events were concentrated in the capitals, starting in the United Kingdom in 2000 and gradually spreading eastwards in 2001 (Poland, Estonia, Lithuania, etc.). We notice the importance of the Luxembourgish and Swiss chapters in terms of number of events and longevity: they both were the most important producers of events in Europe from 2001 to 2005. On the contrary, in the USA the events' count drastically decreased after 2002. In the rest of the world, the presence of the *firsttuesday* events was much more negligible even if they survived until 2004 in South Africa and Australia.

[18] For some fragments, we have not been able to extract a place from the raw text. This is mainly due to an incomplete preservation of the pages.

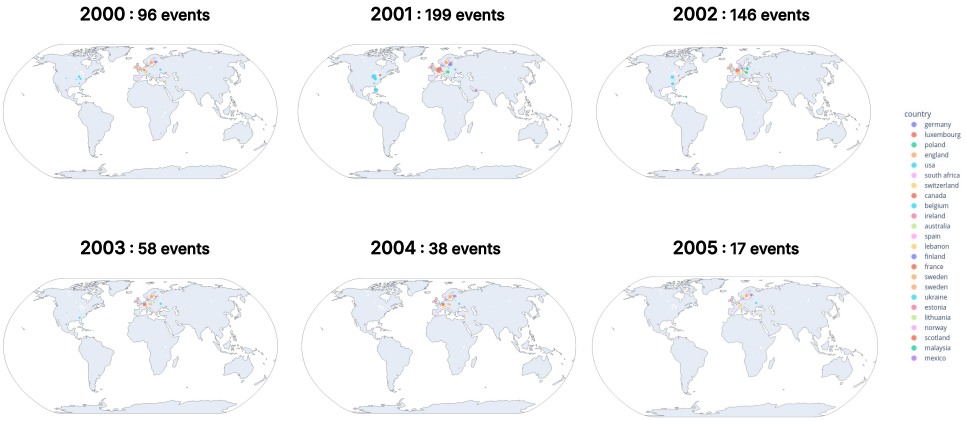

Fig 8. Geographical distribution of the archived *firsttuesday* events per year. Circles are cities. Their size maps the number of events. The colours distinguish the countries.

3.4 Interaction network between speakers and sponsors

Graph theory has long been used by historians to reconstruct evolving networks of social groups from timestamped interactions [46]. In particular, cliques can represent groupings where each entity is related to all others. Cliques might thus reveal precise indications of how an economy is organised by enlightening real and de facto historical associations between social actors. We now use this approach to address the fourth question of historians: *Were there temporally, geographically or thematically structured networks of sponsors and speakers?*

Sponsors played a key role in the development of the *firsttuesday* movement: they made use of their industrial reputation to promote the events, they helped organizers to find prestigious places for the meetings and eventually gave some money. In addition, some speakers (entrepreneurs, startup creators, etc.) were invited to showcase the weekly events and frame the subject of informal discussions that followed. By reconstructing the relationships between those key players through space and time, we might be able to enrich our understanding of the industrial and financial landscape of the *firsttuesday* initiative, to analyse the local audience and organizers. Fortunately, sponsors and speakers were announced within the archived events descriptions. So we first re-dive into the textual content of the 593 events listed previously and, for each event, we extract key players and categorized them as sponsor, speaker or both. We then draw links between every players that participated to a common event. We end up with a complex network of interactions from which we compute cliques to enrich our analysis. Speakers or sponsors who do not belong to any clique are categorised as *isolates*. Speakers or sponsors who belong to at least three different cliques are categorised as *hubs*. The Figure 9 synthesizes our findings.

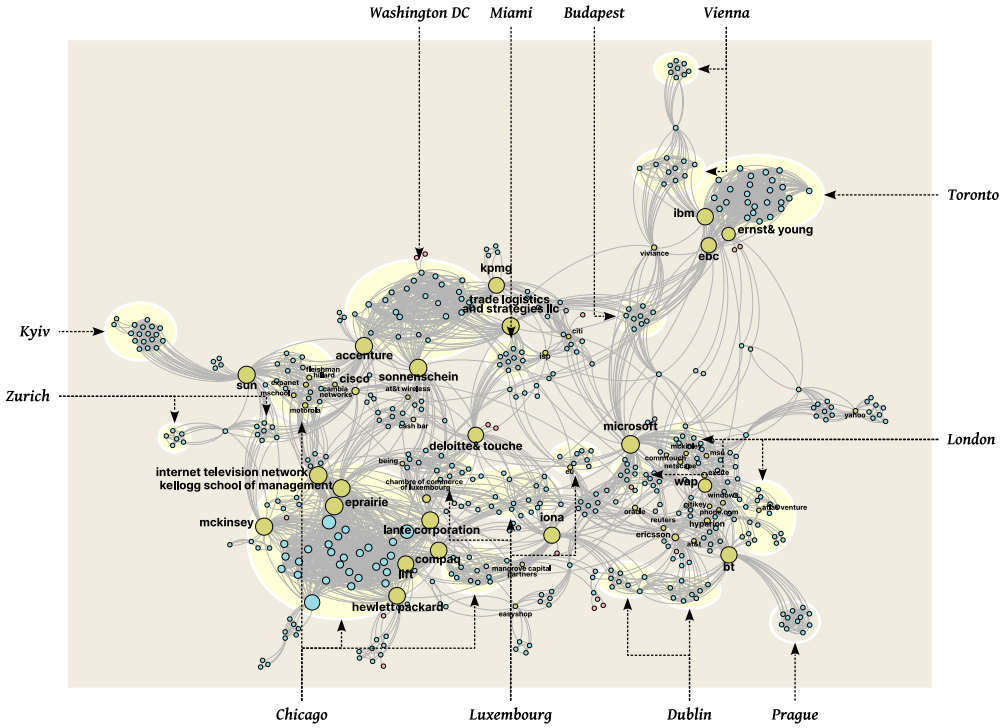

Fig 9. Network of 712 speakers and sponsors. A link indicates that two economic players participated to a common event. Players in blue are part of clique. Players in green are part of at least 3 cliques, there are called hubs. Players in pink are part of no clique. The size of the nodes maps their degree.

This network[19] is made of a single connected component but the unity of the whole is fragile: its structure is only maintained by hubs. All the hubs are in fact sponsors, financial or tech companies involved in the organisation of the events. We thus deduce that the global cohesion of the *firsttuesday* community was not due to the audience or to the invited speakers but to the sponsors and industrial benefactors. We also notice that no sponsor can be categorized as truly global. *Microsoft* and the consulting cabinet *Deloitte* are the only ones with a worldwide dimension as they are somehow central in the network. But the majority of the sponsoring hubs are regional (with two or three local subsidiaries like *IBM*, *Sun*, etc.) or national companies (*Accenture*, *Hewlett-Packard*, etc.). These sponsors are equally distributed between technological actors (electronic or internet-related companies) and financial players (consulting cabinets, banks, law firms, etc.). In Luxembourg and Switerzland, where the domiciliation and administration of investment funds is facilitated by governments, we note the presence of public sponsors such as the Chamber of Commerce of Luxembourg along with private actors. Here, sponsors act as local connectors between unmixed cliques of speakers. This reflects the idea of a developing economic landscape rather than an already established market driven by central players.

The most remarkable feature of this network is the presence of non-overlapping cliques essentially constituted of invited speakers (the dense communities of blues nodes). Non-overlapping cliques are even the most common observable shapes in the Figure 9. Indeed, each clique reflects a geographical and temporal framed group of speakers that didn't moved from one chapter to another. It is thus common to find

[19]Network downloadable at `https://doi.org/10.7910/DVN/EAFOHY`

within a given local chapter two or three unmixed cliques that have only existed for a single year (see the figure 13 in Appendix). We can hypothesise here that the *firsttuesday* community was characterised by a high turnover caused by the collapse of the dot-com bubble and the global instability of the web market. But if all the participants were so isolated within their respective chapters, what then made the movement coherent? Can we still call it a community?

3.5 The evolving semantic landscape of the *firsttuesday* events

We believe that the unity of the *firsttuesday* community is to be found in the vocabulary used to describe the events published online. Through the choice of specific words and expressions or through the themes addressed, symbolic markers and unifying narrative schemes are deployed. We thus now address the last question asked by historian: *Can we study the structure and dynamics of the vocabulary used during the meetings?*

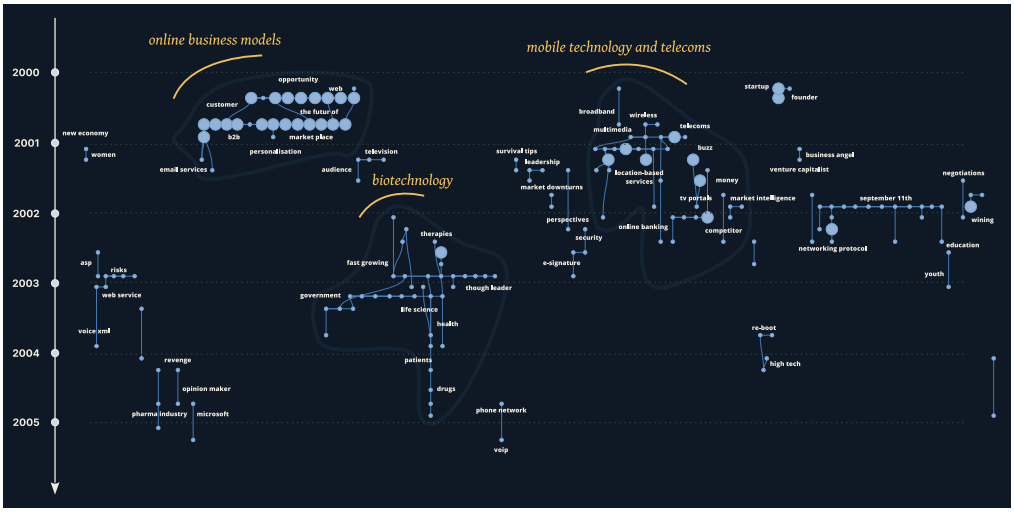

Fig 10. Phylomemy reconstructed from the textual descriptions of 544 *firsttuesday* events. Blue dots are groups of terms or expressions frequently used together. The content of the blue dots can be seen in the interactive version of this phylomemy. White terms are meaningful terms spatialized when and where they appeared in the whole structure. Yellow terms highlight the most important topics discussed during the events.

To that end, we dive into the raw textual content[20] of the events descriptions extracted from the corpus D_4. We acknowledge that these descriptions are incomplete in the sense that they do not reflect the full reality of the offline meetings. But as they were intended to promote the events online, we hypothesize that their textual content was very meaningful, semantically dense and particularly precise. In order to reveal the latent and evolving structure of this vocabulary we choose to use the phylomemy reconstruction method (see 1.6). The corpus of 544 events description is first analysed by GarganText [7] where advanced text-mining algorithms extract a list of 637 meaningful terms and expressions[21]. Then the corpus is divided into fixed periods of 3 months. Within each period, groups of terms frequently used together (the blue circles in the Figure 10) are detected. Then, the phylomemy reconstruction process weaves kinship links between semantically similar groups of terms from one period of time to another. Groups of terms connected through time are called branches. A branch thus

[20] The corpus D_4 is reduced to a set of 544 documents containing raw textual descriptions.

[21] The corpus, the list of terms and the phylomemy can be downloaded at `https://doi.org/10.7910/DVN/EAFOHY`

represents a time-coherent topic discussed during one or more *firsttuesday* events. The resulting phylomemy can be visualized by means of the Figure 10. This structure translate at a global scale the evolution of the semantic landscape of the *firsttuesday* community from 2000 to 2005.

The temporal distribution of the events follow the same patterns than the geospatial analysis of section 3.3 with a peak of meetings in 2001 and 2002. We however found no geographical specificity in the phylomemy which means that the main topics were debated in an undifferentiated way from the USA to Europe.

On the purely semantic side, three axis of analysis wan be followed to understand the phylomemy. First of all, we notice the continuous presence of elements of vocabulary used to qualify the different participants to the meetings: *venture capitalist*, *financier*, *founder*, *analyst*, *entrepreneurs*, etc. Apart from a few ephemeral attempts to diversify themself (*wirelesswednesday*, *matchmaking*, etc.) and despite the collapse of the dot-com bubble, the model of the *firsttuesday* events didn't changed in five years. The same types of players went on meeting offline weeks after weeks.

Secondly, the phylomemy reveals that the events have been focusing on three different topics of discussion (highlighted in yellow in the Figure 10). These topics can be understood as subsequent pivot moments in dawning web industry. The first one, in 2000, dealt with the question of emerging online business models such as *market places*, *b2b* businesses, *merchant* and *sales web sites*, etc. But as theses models were the most affected by the dot-com crash, the community had to quickly move to another interest: the *wireless* and *mobile* trends in 2001. Influenced by the *telecom* and *media* industry, *firsttuesday* events discussed the *opportunities* to expend the *multimedia* sector on mobile devices and to make money from the *location-based possibilities* of the wireless technology. Finally and simultaneously to the decrease of the community in 2002, the discussion topics specialised in biotechnology and pharma-technology and nanotechnology. We also notice at the same time the emergence of *security* and *education* concerns. On the whole, the phylomemy witnesses the transition from the *new technology* to the *high technology*.

Thirdly, a last type of vocabulary acted as connector between the different technological topics addressed during the events. These particular elements of vocabulary fall under scope of the *entrepreneur myth* [47]. Indeed, since the beginning of the 20th century, entrepreneurs are socio-economical players for whom telling his story and his legend becomes a tool of legitimization through the media or through public meetings such as the *firsttuesday*. The continuity of the phylomemy is thus interspersed by first person interviews whose aim was to create and enrich the mythological figure of the entrepreneur. That is, an individual, who will succeed alone, presented as brilliant or visionary and capable of changing the market and even the society through the force of his ideas. Tinged with *social darwinism* [48], these narratives were reinforced by the dot-com crash as witnessed by the use of the semantic field of survival and by emphasizing on individual success despite a context of collective failure. We notice that the use of these elements of story-telling invisible – as a side effect – two important socio-economical players of the web economy: 1) the employees of the startups and tech companies who are never mentioned in our corpus; 2) the governments, administrations and public players who are only considered as obstacles facing the growth of a new business.

4 Discussion

We believe that the visualizations and analyses obtained in section 3 are the most advanced results we can get from the Internet Archive database. These explorations have given us a good understanding of the evolution of a web site that disappeared over

15 years ago. But if we now want to go further, archives alone will not be enough. Indeed, we have reached the point where a return to the field is necessary. From a Computational Social Sciences perspective, the next step will thus be to confront real participants of the *firsttuesday* events to the results of our research and enrich our analysis with the reality of their memories. We could then refine our visualisations in the light of more accurate research questions.

A second possibility for improvement would be to make use of the quality measure introduced in 2 in order to enrich the spatio-temporal scope of our experimental material. Other web archives initiatives may have preserved additional and complementary web areas to those already harvested by the Internet Archive. National libraries may have a better coverage of the different local chapters than the global but un-specialized Internet Archive database. Multiple archives corpora could thus be combined and the possible quality improvement of the resulting set could eventually be validated with our tools.

The current implementation of the *Déndron* library, and in particular the visualization interface of the web cernes, seems to be reaching a display limit around 20.000 archived pages (or around networks of 30 to 40 archived web sites). This is already a good improvement regarding existing works (see 1.6) and in particular the *Anemone* project that was limited to a single archived web site. But if we want to extend our work to the analysis of larger sub-parts of the Web further developments will be necessary. Our next goal will be to scale up our method to the level of hundreds of archived web sites.

But so far, the *persistences* approach and the use of *web cernes* appear to be very efficient for exploration purposes. Indeed, the shape of the web cernes of a given archived web site can be easily compared to the shapes of its neighbours (see 7). The resulting shapes of web cernes satisfyingly catch the eyes of web archives explorers and help them to discovered intriguing sub-area or even missing zones. We thus wonder whether the shapes of web cernes could help us to classify types of dead web sites. Is there something like a typical shape that could help us to detected in the blink of an eye a forum among hundreds of other dead sites?

Finally, our approach is also limited to the only visible content of the archived web sites. Interaction networks and phylomemies are only reconstructed on the basis of informations strictly displayed on screen (raw textual descriptions, dates, places, etc.). For the moment, we can't investigate the pieces of code behind the structure of the web sites, we can't really question the internal technical choices of the targeted web resources [31]. We need to work on this in order to – for instance – reconstruct a phylomemy of elements of code, of JavaScript functions, of designed principles, etc. We could thus reconstruct and analyse the evolution of the technical layer of the Web.

5 Conclusion

By introducing the notions of *persistences* and *continuity spaces* along with the powerful visualization tools of *web cernes*, our study has made it possible to reconstruct the structural and temporal evolution of an archived web site and to extend this reconstruction to its direct neighbourhood. We have demonstrated that our method can guide researchers through the exploration of large corpora of web archives by revealing, thanks to a quality measure, the archived areas where an in-depth exploration can be conducted. Furthermore, this quality measure allows - on mathematical grounds - to qualify the temporal and structural validity of an analysis conducted on archives, to compare several results based on a common corpus and to quantify the possible enrichment of a set of archives via third-party data. Finally, we have illustrated our method by applying it to the historical analysis of the archives of the *firsttuesday.com*

web site, an analysis that will soon be extended to the entire *firsttuesday.com* constellation with the help of historians.

But we also believe that our study can go beyond the simple exploration of web archives. This paper is called to become the first step of a larger study of the **morphogenesis of the Web**. That is to say, starting from the analysis of large corpora of web archives to reconstruct the evolutionary structure of the Web from its origins to the present day, in order to study its temporal dynamics through different levels of complexity and to analyse the socio-technical mechanisms explaining the emergence of online *collective shapes*[22] (communities, controversies, etc.).

The study of morphogenesis is a field of research in the domain of Complex Systems that studies *the set of mechanisms explaining the appearance of structures and controlling their shapes* [49] (shape of dunes, structure of an embryo, organisation of a system of cities, etc.). Knowing a target process, the morphogenesis is interested in the notions of self-organisation, non-equilibrium states, attractors ... Although morphogenesis is classically associated with natural systems, we have every reason to believe that the Web would be a favourable field of study. Indeed, the empirical observations made following the work of F. Ghitalla [50] have long revealed the quasi-organic nature of the Web: its structure is constantly evolving and results from the chaotic effects of a multitude of interacting processes inside or outside the Web (emergence of networks of sites, technological innovations, socio-political shocks). Online shapes emerge and evolve continuously and all the time and we think that persistences, continuity spaces and web cernes can help researchers to reconstruct them from web archives. So in the near future we will have to ask a fundamental question: what is a shape on the Web? A community of Internet users, a viral image and its derivatives, a string of characters, a network of hyperlinks... ? Are there recurring or characteristic shapes? Do they depend on a specific spatio-temporal or socio-technical context? What factors can explain the viability of an online shape? The set of entities that form a community? The structure of their interactions? The nature of the border that separates them from the rest of the network? How can these shapes be characterized dynamically? etc.

Supporting information

S1 Appendix. We use the free text-mining software GarganText [7] to synthesize the scientific landscape of web archiving. We first extract from the *web of Science*, *Scopus*, *Jstor* and *Hal* a corpus of 738 papers metadata (title, date and abstract) matching the query "web archiv*"[23]. GarganText uses natural language processing algorithms to extract meaningful terms and expressions from the raw documents. We manually refine this first list to shape a definitive set of 1080 representative terms. The conditional probability of having one term knowing another is then computed. The resulting map Figure 1 is thus a semantic network that shows scientific topics connected one another by proximity of uses regarding the original papers. There, nodes are topics and links are proximity measure, this graph is spatialized with *Gephi*. At a global scale, the map reveals communities of topics (numbered dense areas of nodes) that can be understood as sub research fields of web archiving.

S2 Appendix. Déndron is a free (AGPL + CECILL v3) Python library (Python V3) for exploring web archives. Déndron is made of 9 python scripts. An example command is added at the beginning of each script. The scripts produce one or more output file

[22] Collective shapes are shapes resulting from entities in dynamic interactions within a given environment [49]

[23] The corpus, the vocabulary can be download at `https://doi.org/10.7910/DVN/EAFOHY`

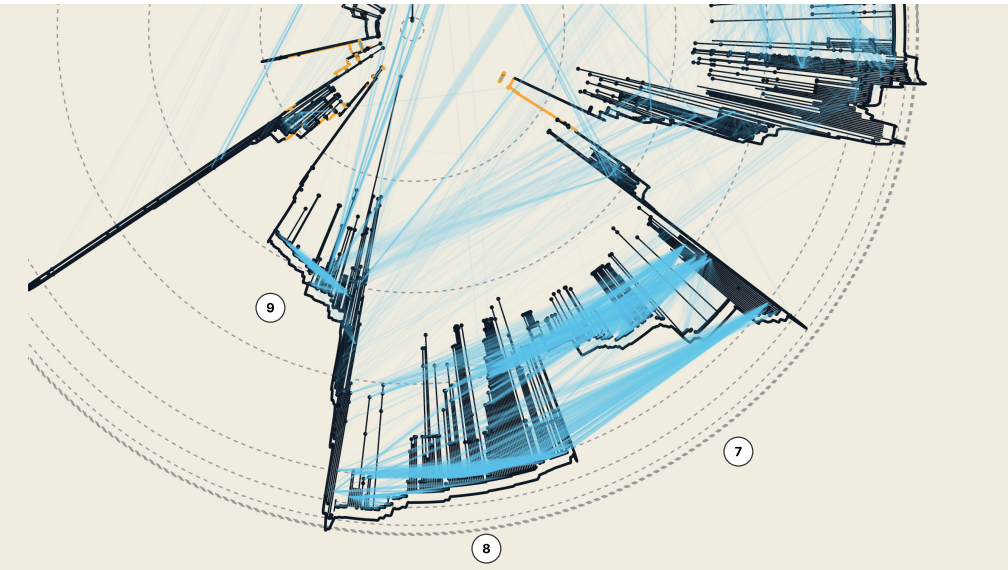

Fig 11. Close up of the web cernes reconstructed from the archives of *firsttuesday.com*. We here focus on partitions 7,8 and 9, aka the events' descriptions. The blue lines features continuity cliques between persistences (the black lines).

saved in `/output/`. `getArchives.py` requests for all the unique archived urls sharing the same prefix from the Internet Archives API. `archivesToLinks.py` gets all the internal/external hyperlinks from a list of archived urls. `sortLinks.py` distributes raw archived citation links among internals, externals, neighbourhood and emails dedicated lists. `linksToUnarchived.py` finds unarchived pages among internal archived links. `linksToSiteTree.py` reconstructs the tree like structure of a web site from its archived and unarchived pages. `archivesToPersistences.py` groups the archived and unarchived pages by url and aggregate their timestamped snapshots to create persistences. `persistencesToGraph.py` finds continuity links between persistences. `graphToCliques.py` finds cliques within a continuity graph. `graphToCernes.py` sorts the archived an unarchived pages regarding their continuity links and href citations, it then saves a file of web cernes. In addition, in the folder `/viz/`, there is a standalone web page `web-cernes.html` designed to visualize and explore the web cernes produced by the script `graphToCernes.py`. Déndron is downloadable at `https://gitlab.iscpif.fr/qlobbe/dendron`.

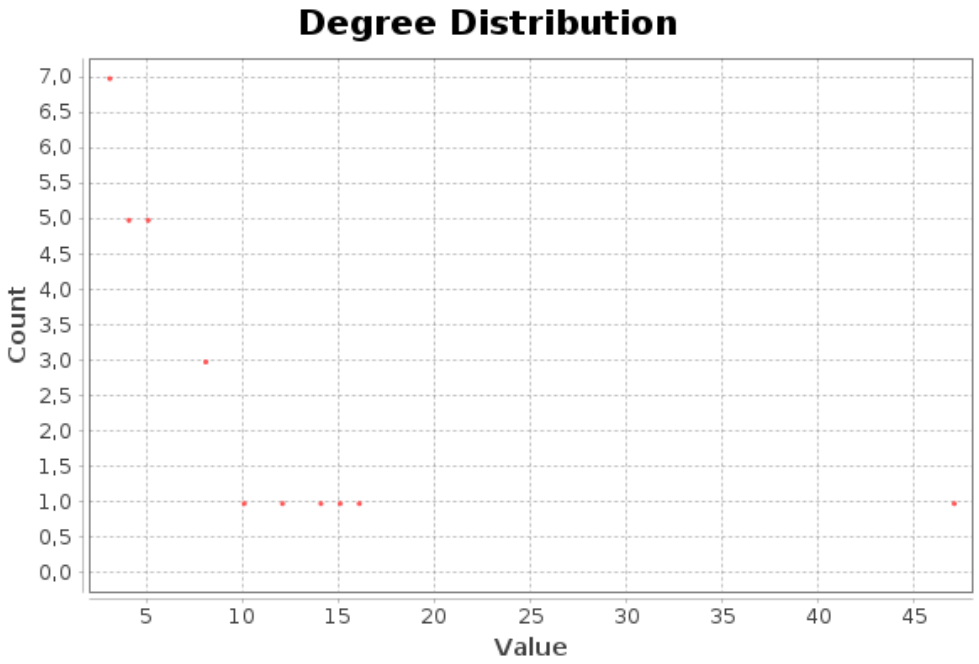

Fig 12. Degree distribution computed with Gephi of the *firsttuesday* constellation

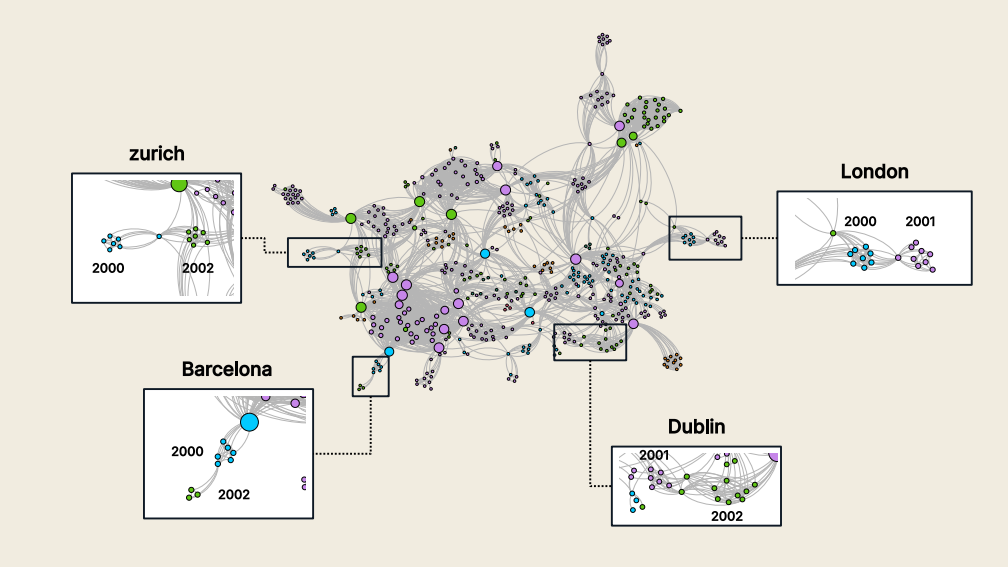

Fig 13. Close up of chapters made of non-overlapping cliques but related to a specific year. Nodes are colors by years (blue for 2000, purple for 2001, light green for 2002, orange for 2003, dark green for 2004 and pink for 2005)

References

1. UNESCO. Charter on the Preservation of Digital Heritage; 2003.
2. Masanes J. Web archiving: issues and methods. In: Web archiving. Springer; 2006. p. 1–53.
3. Lobbé Q. Where the dead blogs are. In: International Conference on Asian Digital Libraries. Springer; 2018. p. 112–123.
4. Flichy* P. Genèse du discours sur la nouvelle économie aux États-Unis. Revue économique. 2001;52(7):379–399.
5. Kahle B. Preserving the internet. Scientific American. 1997;276(3):82–83.
6. CERN. The document that officially put the World Wide Web into the public domain; 1993. Available from: `http://cds.cern.ch/record/1164399`.
7. Delanoë A, Chavalarias D. GarganText, collaborative decentralized software; 2023. `https://gitlab.iscpif.fr/gargantext/main`.
8. Schafer V, Thierry BG. The "Web of pros" in the 1990s: The professional acclimation of the World Wide Web in France. New Media & Society. 2016;18(7):1143–1158.
9. Ben-David A, Amram A, Bekkerman R. The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. International Journal on Digital Libraries. 2018;19(1):95–106.
10. Gebeil S. Web archives as sources for the uses of the past during the 2000s. For a social, micro-historical and qualitative approach. In: Audiovisuals and internet archives: Histories of healthy bodies in the 21st century ERC Body Capital Spring School 1-4 April 2019; 2019.
11. Chavalarias D, Lobbé Q, Delanoë A. Draw me Science – multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. Scientometrics. 2021;doi:10.1007/s11192-021-04186-5.
12. Lobbé Q, Chavalarias D, Delanoë A. On Level and Scale in Digital Humanities. In: Armaselu F, Fickers A, editors. Zoomland. Exploring Scale in Digital History and Humanities. De Gruyter; 2023 (to be published).
13. Spaniol M, Denev D, Mazeika A, Weikum G, Senellart P. Data quality in web archiving. In: Proceedings of the 3rd Workshop on Information Credibility on the Web; 2009. p. 19–26.
14. Cai D, Yu S, Wen JR, Ma WY. Extracting content structure for web pages based on visual representation. In: Asia-Pacific Web Conference. Springer; 2003. p. 406–417.
15. Song J, Wang D, Bao Y, Shen D. Collecting and storing Web archive based on page block. Journal of Software. 2008;19(2):275–292.
16. Sanoja A, Gançarski S. Yet another hybrid segmentation tool. In: iPRES 2012–9 th International Conference on Preservation of Digital Objects; 2012.
17. Oita M, Senellart P. FOREST: Focused Object Retrieval by Exploiting Significant Tag Paths. In: WebDB. Melbourne, Australia; 2015. p. 55–61. Available from: `https://hal-imt.archives-ouvertes.fr/hal-01178402`.

18. Brügger N. Website history and the website as an object of study. New Media & Society. 2009;11(1-2):115–132.

19. Mackinnon K. The death of GeoCities: seeking destruction and platform eulogies in Web archives. Internet Histories. 2022; p. 1–16.

20. AlSum A. Reconstruction of the US First Website. In: Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries; 2015. p. 285–286.

21. Jatowt A, Kawai Y, Tanaka K. Visualizing historical content of web pages. In: Proceedings of the 17th international conference on World Wide Web; 2008. p. 1221–1222.

22. Toyoda M, Kitsuregawa M. A system for visualizing and analyzing the evolution of the web with a time series of graphs. In: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia; 2005. p. 151–160.

23. Raffal H. Tracing the online development of the Ministry of Defence and Armed Forces through the UK web archive. Internet Histories. 2018;2(1-2):156–178.

24. Brügger N. Tracing a historical development of conspiracy theory networks on the web: The hyperlink network of vaccine hesitancy on the Danish web 2006–2015. Convergence. 2022; p. 13548565221104989.

25. Weltevrede E, Helmond A. Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. First Monday. 2012;.

26. Hale SA, Yasseri T, Cowls J, Meyer ET, Schroeder R, Margetts H. Mapping the UK webspace: fifteen years of British universities on the web. In: Proceedings of the 2014 ACM conference on Web science; 2014. p. 62–70.

27. Holzmann H, Nejdl W, Anand A. The Dawn of today's popular domains: A study of the archived German Web over 18 years. In: 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL). IEEE; 2016. p. 73–82.

28. Ben-David A. What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. New Media & Society. 2016;18(7):1103–1119.

29. Cocciolo A. The rise and fall of text on the Web: a quantitative study of Web archives. Information Research: An International Electronic Journal. 2015;20(3):n3.

30. Schafer V, Truc G, Badouard R, Castex L, Musiani F. Paris and Nice terrorist attacks: Exploring Twitter and web archives. Media, War & Conflict. 2019;12(2):153–170.

31. Nielsen J. Experimenting with computational methods for large-scale studies of tracking technologies in web archives. Internet Histories. 2019;3(3-4):293–315.

32. Fry BJ. Organic information design. Massachusetts Institute of Technology; 2000.

33. Chavalarias D, Lobbé Q, Delanoë A. Draw me Science – multi-level and multi-scale reconstruction of knowledge dynamics with phylomemies. Scientometrics. 2021;doi:10.1007/s11192-021-04186-5.

34. Lobbé Q, Delanoë A, Chavalarias D. Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. Information Visualization. 2021; p. 14738716211044829.

35. Abélès M. Nouveaux Riches (Les): Un ethnologue dans la Silicon Valley. Odile Jacob; 2002.

36. Flecher M. Le monde des start-up: le nouveau visage du capitalisme? Enquête sur les modes de création et d'organisation des start-up en France et aux États-Unis. Université Paris sciences et lettres; 2021.

37. Delanoë A, Chavalarias D. Mining the digital society - Gargantext, a macroscope for collaborative analysis and exploration of textual corpora. 2023 (upcoming);.

38. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the international AAAI conference on web and social media. vol. 3; 2009. p. 361–362.

39. Diminescu D. e-Diasporas Atlas: Exploration and cartography of diasporas on digital networks. Ed. de la Maison des sciences de l'homme; 2012.

40. Bergson H. Durée et simultanéité: à propos de la théorie d'Einstein. F. Alcan; 1926.

41. Toyoda M, Kitsuregawa M. What's really new on the web? identifying new pages from a series of unstable web snapshots. In: Proceedings of the 15th international conference on World Wide Web; 2006. p. 233–241.

42. Saad MB, Gançarski S. Improving the quality of web archives through the importance of changes. In: Database and Expert Systems Applications: 22nd International Conference, DEXA 2011, Toulouse, France, August 29-September 2, 2011. Proceedings, Part I 22. Springer; 2011. p. 394–409.

43. Wills P, Meyer FG. Metrics for graph comparison: a practitioner's guide. Plos one. 2020;15(2):e0228728.

44. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment. 2008;2008(10):P10008.

45. Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, et al. Graph structure in the web. Computer networks. 2000;33(1-6):309–320.

46. Gardin JC, Garelli P. Étude par ordinateurs des établissements assyriens en Cappadoce. Annales. 1961;16(5):837–876.

47. Galluzzo A. Le mythe de l'entrepreneur: défaire l'imaginaire de la Silicon Valley; 2023.

48. Rogers JA. Darwinism and Social Darwinism. Journal of the History of Ideas. 1972;33(2):265–280.

49. Bourgine P, Lesne A. Morphogenèse. L'origine des formes: L'origine des formes. Belin Éducation; 2015.

50. Ghitalla F. Qu'est-ce que la cartographie du web ? Boullier D, Jacomy M, editors. OpenEdition Press; 2021. Available from: https://doi.org/10.4000/books.oep.15358.