New Media Research Seminar: Assignment 3

Digital Methods and the Internet Archive

Textual Grammar and Website History

What narratives emerge from the history of a website? What Niels Brügger calls a site's "textual grammar," from its structure and features to metaphors used for navigation, may reflect not only changing Web aesthetics but also broader developments both on- and offline. The video Google and the Politics of Tabs tracks the directory's disappearance from Google's home page, but in so doing makes a larger point about expert ordering of information being displaced by the algorithm. A sequel could perhaps feature Yahoo and Technorati, where attempts have been made to remediate the directory in light of the engine. Similarly, one might consider the history of YouTube when discussing the origins of social software: there, one sees a shift from a focus on dating and personal expression to a service more compatible with traditional content production. Keep in mind that website history is not necessarily web history. What story is told, for example, by the elaboration and professionalization of the 9/11 conspiracy website, 911truth.org?

Assignment: Capture past versions of a website via the Internet Archive, outputting a video with narration. Consult also Digital Methods, 'The Website,'
(http://www.digitalmethods.net/Digitalmethods/TheWebsite) and the 13 step guide provided below.

As for the assignment for next week, these are the (technical) steps to create a movie from the Wayback archive:
1. Type http://www.google.com into the Internet Archive Wayback Machine
Link Ripper
(http://tools.issuecrawler.net/beta/internetArchiveWaybackMachineLinkRipper/)
2. Right-click and save the text-file to your computer

3. We assume you have installed Firefox 3.x (http://getfirefox.com) as well as the Firefox extension Grab Them All (https://addons.mozilla.org/en-US/firefox/addon/7800). Here is a screencast on how to install Grab Them All (http://rafal.zelazko.info/2008/06/23/easy-screenshot-of-many-sites/)
4. In Firefox, go to Tools > Add-ons. Make sure the preferences of Grab Them All are as follows: PNG, Grab visible window only, Safe URL, generate report file, width: 1024, height: 768.
5. In Firefox, go to Tools > Grab Them All.
6. Click on 'Load file with URLs to grab' and point it to the text file you just downloaded.
7. Select a directory where all the screenshots should be saved.
8. Set the 'Max processing time' to 20 or 30 (the Wayback Machine is very slow).
9. Set 'Javascript time to wait' to 10.
10. Click "Let's go!".
11.1 Go to bed and check your results in the morning.
11.2 See if all went well by looking at the pictures. Sometimes the archive returns an error. Note those URLs in a new text file and repeat steps 4 to 9 until all screenshots succeeded.
12. Load png's into an image viewer like e.g. iPhoto.
13. Make project into iMovie, sorted by name, which keeps the pages in chronological order. Record Narrative

OR

Go to http://screenr.com and record your narrated slideshow through this web service

Michael Stevenson

The Archived Blogosphere, 1999-2001

Changing Cultures: Cultures of Change
Draft conference paper
November 2009

The Internet Archive (www.archive.org) is perhaps the most important and certainly the largest resource available to the web historian. Currently it exceeds 2 perabytes of data, with archived versions of web pages dating back to 1996. The archive has proved useful in a variety of contexts, whether providing webmasters with earlier versions of their site or acting as evidence in criminal investigations and in court. As a resource for studying past states of the Web, however, the archive's potential has yet to be fully explored. The archive's interface, the Wayback Machine, makes it easy to consult a previous version of a web page, and can also relatively easily be used to track the evolution of a particular page or site over time. There are also special Web collections - expert lists of sites that converge on a particular topic or event. Beginning with a collection covering the 1996 U.S. presidential elections, the practice of creating special collections around momentous world events has only grown, with for example the creation of the September 11 Web Archive in the hours following the World Trade Center attacks, as well as the Asian Tsunami Web Archive in December 2004. If archives are the material with which history is constructed (Lynch, 1999), then one might further argue that the means by which one can access the archive may privilege certain kinds of history writing. Two of the most familiar approaches, one biographical and the other the history of events, are not only reflected in single site histories and special collections, but are possibly privileged to the extent that methods of collection and categorization take them for granted. Attention to the constructedness of archives may not be new, but the specificity with which the Internet Archive was created is nonetheless important. Beyond considering limits to current methods, however, the question is how else the archive as currently deployed may be navigated and put to use. Here, I will discuss a case study in which researchers and I looked to innovate historical web research by applying insights from Digital Methods Initiative (www.digitalmethods.net) to the archived blogosphere.

The Internet Archive as Legacy System

Before turning to the case study, however, it is worth recounting some of the challenges involved in archiving the Web as well as resulting issues for the web historian. First and foremost, of course, is the issue of sheer quantity - as one journalist noted in 2005, despite some 40 billion pages, the Internet Archive covers only a "fraction" of the web's history (Boutin, 2005). The archive's index (taken from Alexa.com) would seem to benefit the most popular sites at the expense of less frequently visited sites and the so-called "dark web," or sites that lack sufficient inlinks. But the problem of incompleteness is not limited to issues of indexing and hard drive space: there are also, for instance, blocked sites (the Internet Archive's crawlers, for instance, obey robots.txt), outdated software and formats that resist preservation. The Internet Archive's founder, Brewster Kahle, notes that Flash and complex Javascript files are especially problematic (Rein, 2004). The difficulty involved in archiving the web is made clear through comparison with other media. The web, Steve Schneider and Kirsten Foot write, is "a unique mixture of the ephemeral and the permanent" (2004: 115), meaning it is in part comparable to live media such as radio and theater and in other ways to permanent media such as print and film. Preservation of the former requires some act of recording or mediation, while the latter are generally preserved in their original form. The web, however, does not fall neatly into either category. In order for a website to be transmitted, it must have a permanent or semi-permanent home on a server; at the same time, "a website may destroy its predecessor regularly and procedurally each time it is updated by its producer" (ibid). With this in mind, Brügger argues that any archived website is a "subjective reconstruction" (2009: 125), as it requires not only choices of what to include and what to omit, but the timing of the archiving process will also 'shape' the object of study. For example, an image linked to and shown on a page may change while other contents stay the same. This ephemerality helps explain why the Internet Archive's contents are often referred to as "snapshots" of the web (e.g. Kahle, 1996; Internet Archive, "About").

These snapshots can be accessed through the Wayback Machine, where users are prompted to input a URL. If the page has been archived, the query returns a list of dates, each standing for a timestamped version of the URL requested - the list may include dates annotated with an asterisk, indicating page updates. In the advanced search, one is additionally given the options to filter by date or file type, to use wild cards for the URL (thus potentially returning all pages from a particular domain) and to use third-party services to either compare two versions of a page or to create a pdf of the page. The Wayback Machine may also serve one of a few error messages, including "Not in

Archive," which states that no version exists and the URL will be crawled at a later time, "Failed Connection," which means the page is in the archive but the server on which it is hosted is down, and "Blocked Site," displayed when a URL has been indexed but has either not been crawled or has been removed from the archive, because the owner has disallowed the archive's crawler via Robots.txt. Generally, however, clicking on the desired date serves the archived page. From there, owing to an innovative feature in the Wayback Machine, users can navigate the archive by surfing: clicking on a link takes her to the version of that page closest to the source of the link in terms of date. But in line with Brügger's concerns about the convoluted temporality of the archive, it's worth noting that such navigation always requires a "jump-cut" through time, sometimes a matter of a few minutes but often one of considerable length. Clicking a link to the Arts and Humanities index from Yahoo!'s homepage in 2000, for instance, will take the user to a version archived in 2002. When the page requested is not in the archive, users are returned to the live Web.

For those first becoming acquainted with the archive, one question seems unavoidable: how do you search? One cannot query the archive for keywords - instead, one begins with a location, a form of navigation that may seem jarring to users more comfortable with search boxes than address bars. From this point of view, one could go so far as to say the archive and its Wayback Machine interface are examples of a 'legacy system' - technology that has been superseded, yet remains in use because it is too difficult, expensive or cumbersome to replace - not to mention the fact that we cannot go 'wayback' and change the way we archive. If that were possible, one could not only imagine an alternative archive that allows search, but perhaps one that also saves contextual information, such as a site's estimated number of visitors and its placement in search returns from the corresponding time period. In this way, an improved archive might document the rise and fall of sources of information for a particular query, similar to the Issue Dramaturg created by Govcom.org. But the internet archive may also be considered a legacy system in another sense, one more at home in the humanities than in computer science, and one that may go some way to explaining why the archive was given the particular form it has. In this other sense, a legacy system sustains aspects of an earlier (web) culture, one that may seem more or less forgotten but nonetheless persists discursively and - perhaps more importantly - materially as a format for communication, navigation, participation and interaction on the Web. That is to say, the legacies of a past web culture may be studied and assessed through attention to the forms taken on by contemporary web systems.

In this sense of the legacy system, what is sustained by the Internet Archive and its Wayback interface is cyberculture. Cyberculture is itself not a straightforward concept, but refers in general to

the culture of the Internet and early Web, especially in the 1980s and 1990s, a period most often characterized by the great deal of excitement and hype surrounding the new technology (see Silver, 2000, on "popular cyberculture"). As a broad starting point, one might think of cyberculture as a kind of technological exceptionalism - the belief that this medium would be different. In a 1996 article that acts as the mission statement for the Internet archive, Kahle wrote that the library would have to be reinvented in order to preserve the Internet. And while he assumed such an archive would eventually be used in ways remarkable and remarkably different to previous libraries, these applications were not yet visible when the archive was created (Kahle, 1996). The future uses of the archive would exacerbate the already radical progress represented by the new medium: this is perhaps why, when Kahle uses the term library to describe the work of the Internet Archive, it appears in scare quotes (ibid).

Related to this is cyberculture's commitment to egalitarian and universal access to information. In addition to the sense of a notional space without borders, cyberculture encompasses the belief that the public availability of information is central to progress and innovation. This given, as others have noted, reflects not only a dedication to Enlightenment principles, but is often accompanied by a level of zeal that borders on evangelical. Kahle, who also sits on the board of the Electronic Freedom Foundation, and whose company *Alexa* was named after the Library of Alexandria, is a self-avowed "silicon-valley, utopian" type, unsurprising considering the scope and scale of his ambitions (Kahle, 2007). Critics of the "Californian ideology" have long accused Web entrepreneurs of naive technological determinism as well as a blindness to issues of social and economic inequality that are exacerbated rather than solved by information technologies (Barbrook and Cameron, 1996), but it is also possible to see in the rhetoric and actions of Kahle and other Web "gurus" more subtle beliefs and assumptions embedded in cyberculture. One of these is what could be called the primacy of information, or to borrow from N. Katherine Hayles, the perception of information as "more mobile, more important, more essential than material forms" (1999: 19). In other words, culture is reduced to quantifiable yet itinerant data, separated from its material instantiation (and often its historical and geographical context). This is at work, for instance, when Kahle considers the archive's size relative to that of other cultural expressions.

- A video store holds about 5,000 video titles, or about 7 terabytes of compressed data.
- A music radio station holds about 10,000 LP's and CD's or about 5 terabytes of uncompressed data.
- The Library of Congress contain about 20 million volumes, or about 20 terabytes text if typed into a computer.
- A semester of classroom lectures of a small college is about 18 terabytes of compressed data (Kahle, 1996).

In addition to serving as an example of an informational sublime in the strand of Gibson's cyberspace (in terms of size, the Internet Archive has long surpassed the figures Kahle quotes), such offhand references to various cultural processes as different organizations of data obviously neutralize their production. As Hayles goes on to point out, "When this impression [of the primacy of information] becomes part of your cultural mindset, you have entered the condition of virtuality" (1999:19).

Finally, there is cyberspace itself. As Wendy Chun (2006) and others have argued, cyberspace as a metaphor for the Web is non-sensical - cybernetics, or the science of communication and control, has actually very little to do with space, but is characterized by the concepts of negative feedback, the servomechanism and homeostasis. Nonetheless, with its popular connotations of a virtual existence and a potential for excitement and discovery on the "electronic frontier," cyberspace came to represent a number of conventions regarding the supposed experience of being online, if not how the Web actually worked. Early browsers such as the Netscape Navigator, Internet Explorer and Safari played off the notion of an unwieldy information space (Rogers, 2009), while cyberspace and virtual worlds were centerpieces in stories of experimental identity and the primacy of the Symbolic over the Real (Dibbell, 199?). In what is becoming archaic language, the cybernaut of science fiction and the early Web would "surf" from site to site, at most making use of a directory for more methodical browsing. With the rise of search engines, always-on connections and Web 2.0 platforms, this era appears to have passed. In a *Newsweek* article on the services collectively known as "Web 2.0," Steven Levy and Brad Stone (2006) summarized the shift as follows:

> Less than a decade ago, when we were first getting used to the idea of an Internet, people described the act of going online as venturing into some foreign realm called cyberspace. But that metaphor no longer applies. MySpace, Flickr and all the other newcomers aren't places to go, but things to do, ways to express yourself, means to connect with others and extend your own horizons. Cyberspace was somewhere else. The Web is where we live.

In both the Wayback Machine and the archive's special collections, however, "somewhere else" remains visible. This extends beyond the archive's contents, which include a collection on Web Pioneers, and into the formal aspects of the archive's interface. With its emphasis on location rather than keyword, on sites as opposed to today's platforms, and directory-like special collections rather than search, and inadvertently through its failed connections and lag times, the archive preserves not only the historical documents that constitute the early Web but also the look and feel of cyberspace. The archive's preference for the : asked how to ensure a site is archived, Kahle says that webmasters should strive to "keep things fairly simple" and make "more straightforward use of

pointers, because the hyperlink is one of the great ideas of the Internet" (Rein, 2004). Invoking once again the typical Google user of today's Web, the act of consulting the Wayback machine may itself seem like a journey into the past.

Formatting the Archive: Single Site Histories and Special Collections

While the Wayback machine serves primarily the purposes of nostalgia (Kahle, 2007), its higher-profile uses extend to its status as legal evidence and as a means for preserving cultural heritage, particularly through event-themed special collections. Web forensics describes the use of web archives in criminal investigations and legal contexts, paradigmatically in the case *Telewisja Polska USA, Inc v. Echostar Satellite Corporation* (Howell, 2006). The case hinged on whether the Polish television channel advertised its inclusion on the Echostar subscription service after a contract between the two parties expired. A date-stamped version of the Polska's website showed this was the case, and here, as in other lawsuits, evidence taken from the Internet Archive was admitted (ibid).

The status of archived websites as evidence is extended in historical research. Web archiving scholar Niels Brügger outlines the work of website history as follows:

> One of the tasks that could be undertaken would be to formulate a historical textual grammar of the website [...] to ask how the textual elements that constitute the website, along with the semantic, formal and physically performative relations between them, actually have appeared and functioned at various periods in the past in order to identify recurrent patterns and traits (2009: 128).

Evidence of a shifting role or status of a particular website, in other words, will appear in the changes made to it over time. Brügger's ongoing research project DRDK.DK , on the first ten years of Danmark Radio Online, takes the website as an entry point to studying the recent history of Danish media and the history of the Danish Web.[1]

The work of distilling the "textual grammar" of a website's history is demonstrated in "Google and the Politics of Tabs," a video from the Digital Methods Initiative (DMI) that explores changes made to Google's homepage over an six-year period. The video tells the story of the decline of the human-edited web vis-a-vis the search engine algorithm. From October 2001 Google placed the DMOZ directory prominently on its front page, first in the center of the page and later as one of a number of tabs that served as alternative entry points to the Web (including Images, Groups and News), until the directory was relegated to a "more" page in March 2003, and later to a page

---

[1] See DRDK.DK: dr.dks historie 1996-2006, available at http://www.drdk.dk (accessed 29 October 2009).

labelled “even more.” In August 2007, the directory was no longer accessible from Google’s front page - to find the directory, one would have to search.

As in Brügger’s work, DMI focuses on the history of a single site. However, where Brügger’s approach assumes the website as a central, if not natural, object of study in Internet history, DMI makes a point of asking what types of methods are privileged by dominant Web devices. For example, to “delimit” the blog (I’m using Brügger’s term for establishing the object of study), one would ask how the blog search engines - Technorati and Google blog search - already do that work: a defining element for Technorati, for example, is the RSS feed. In other words, following the medium, one could argue that the object of study is constructed by the devices that govern it. Applying this to the Wayback Machine, one notes that the core unit is the host url and the pages associated with it, and that the method of historical analysis privileged (for example by the asterisk denoting updates to a page) is to track a particular site’s evolution. “In effect, the Internet Archive, through the interface of the Wayback Machine, has organized the story of the Web into the histories of single Websites” (Digital Methods Initiative, “The Website”).

In terms of approaches to archive collection and presentation, one alternative to the history of single-sites provided by the Wayback Machine is the specialized collection, which range from archives devoted to the preservation of a particular source or type of website to what are called thematic collections. The gloomily named CyberCemetery, hosted by the University of North Texas, preserves the websites of government agencies and commissions that have been discontinued, while thematic collections include the September 11 Web Archive and the 2004 Asian Tsunami Web Archive.[2] These latter collections, as well as others that were created for various U.S. elections since 1992, are examples of “thematic crawling,” as opposed to “broad based crawling” - the former takes as its starting points sites relevant to a particular theme, while the latter ‘snowballs’ from sites not necessarily related, indexing and capturing new sites through hyperlinks (Schneider et al., 2003). Thematic crawling attempts to exhaustively preserve a pre-defined set of websites that are related to a particular theme (though in some cases the list will be edited and expanded over time), whereas broad-based crawls such as those on which the Internet Archive relies continually branch out to incorporate new web objects. Additionally, due to their smaller scope, the special collections in some cases can be searched, and have generally been complemented with manually-added catalog data: the September 11 Web Archive, for instance, can be browsed by producer name, type and country, as well as according to a limited number of Library of Congress subjects. Broadly

---

[2] Available at http://govinfo.library.unt.edu, http://september11.archive.org and tsunami.archive.org, respectively (all accessed 30 October 2009).

considered, the thematic collection presents world events as they were covered and enacted on the Web. If the Wayback Machine organizes Web history as the evolution of single sites, then these collections reveal this history as a series of momentous occasions as seen through a variety of actors on the Web.

Detailing the affordances and constraints of thematic collections, the September 11 Web Archive team points out that the where the theme-based crawl allows for more certainty that the archived materials accurately represent the sites as they existed, there is a greater risk of previously unrecognized sources being left out (Schneider et al., 2003). The theme-based crawl may therefore preclude what the authors consider a more ideal method of collection and analysis, described under the heading of the Web sphere (ibid; Schneider and Foot, 2004). The Web sphere is defined as "not simply a collection of websites, but as a hyperlinked set of dynamically-defined digital resources that span multiple websites and are deemed relevant, or related, to a central theme or 'object'" (Schneider and Foot, 2004: 118). By "dynamically-defined," the authors refer to changes in the collection over time, particularly with the appearance of new sources within a hyperlinked network (ibid; Foot and Schneider, 2002). Ideally, then, a web sphere approach would combine the focus of a thematic collection with the flexibility provided by a broad-based crawl.

The question of how a web archive is collected and formatted is tied to perceptions of both the structure of the Web and of Web history. With the Internet Archive and the various special collections, web historians are limited to querying URLs in the Wayback Machine or browsing and searching thematic collections. The challenge for those looking to research the early web while resisting single-site and event-based historiographies is to develop methods that engage with the broad-based archive, but offer alternative forms of organizing and analyzing its contents.

The Archived Blogosphere

"Cyberspace was somewhere else. The Web is where we live" (Levy and Stone, 2006). The various pronouncements of the Web's rebirth as '2.0' were built on the rejection of previous visions of the Web and its significance. In particular, the virtual existence implied by the concept of cyberspace appeared to lose currency as a description of the Internet's promise. Arguably, one early source of this rejection of cyberspace was the blogosphere as it emerged around 1999. From the beginning, the weblog was distinguished from the 'site' as well as formats imported from older media, including the news page and instead was understood as a running commentary, in an early definition literally a log of sites visited (Barger, 1999b). Later accounts would highlight the

personal 'voice' of blogs, as well as their engagement with traditional media through links and commentary (Blood, 2000; 2002). Proponents consider these characteristics, alongside the now-familiar reverse-chronological order of posts, as evidence of the blog being "native" to the Web (Blood, 2002), an expression of the Web's formal DNA (Rosenberg, 2009). Emphasis was also put on their conversational quality: a group of individual but interlinked blogs would avoid the "flame-wars" and other tragedies of the commons that routinely undermined message boards and virtual communities (Katz, [199?]2002). The term blogosphere was originally introduced as a joke in 1999, a tongue-in-cheek response to the media attention given at the time to a number of bloggers - tellingly, however, it was coined as a replacement for a previous hype, that of "cyberspace" (Graham, 1999).

The rise of the blogosphere, then, can perhaps be seen as a marker within the larger transition from a period of early Web history, or cyberculture, to what might today be called Web culture. With blogging, this shift was manifested discursively and in Web practices, with the establishment and popularization of the reverse-chronological or "real-time" presentation, an engagement with traditional news media and a signature ironic writing style, called 'snark.' Blogging also represented (if not created) a more popular perception of the value of links beyond navigational tools, namely as measures of association and reputation (though this development was of course also key for search engines). Bloggers - especially Jorn Barger - would discuss at length the importance of quality linking, and promoted the idea that the Web could and would be filtered in an organic, democratic process (Rosenberg, 2009; Barger, 1999a).

The link-plus-commentary style of blogging had two other, as yet under-interrogated, effects in the creation of the blogosphere. The first deals with the effects of the blog's status as miscellany (Dibbell, 2002). That is, the blogosphere would be inclusive through its tendency, for instance, to link to valuable sources outside of mainstream news, or by combining popular and alternative culture (Blood, 2000). Following individual preferences rather than any formal guidelines, a blog's links would reveal ultimately reveal "personality" as opposed to ideology (ibid). Second, in emphasizing links as indicators of reputation, bloggers, alongside more conventional publicity, helped create what would be called the 'A-List,' or the set of bloggers perceived to have disproportionate influence and presence. Both aspects were central to understandings of the early blogosphere, if the anthology *We've Got Blog* (Rodzilla, ed., 2002) is any indication, where a number of posts and articles characterize the blogosphere as either a source of a critical media sensibility or one of media hype and celebrity (and sometimes both).

If the early blogosphere should be seen as the beginnings of a post-cyberspace Web, how does this affect efforts to map it in the Internet Archive, argued above to be a cybercultural legacy system? In addition to the seminal archiving issues of selection and incompleteness, as well as those specific to the Web (sites that disallow crawling, or use software that prohibits archiving), capturing an interlinked sphere within the Internet Archive requires a consideration of starting points - the question of which blogs to include and what periods to study - as well as a means to display the interconnections and associations that constitute the blogosphere - that is, to establish context for the archived blogs. The case study presented here takes up the following questions: What entry point for studying the early blogosphere; what portion can be accessed through the Internet Archive? How to locate the A-list blogs, and how to characterize or 'profile' the early blogosphere in terms of its associations, or the sources most commonly linked to? Finally, how to map the blogosphere, showing clustering by type of blog or around particular sources, and track changes over time?

What entry point for studying the early blogosphere; what portion can be accessed through the Internet Archive?

The EatonWeb Portal was created in 1999 by Brigitte Eaton, and was the first blog directory. Ito originally included some 30 blogs, a number that grew to 400 by November 1999. By that time, Eaton had given up trying to track the appearance of new blogs on her own, and included a form for readers to anonymously suggest blogs. On August 15, 2000 - the date of the earliest complete version of the directory available in the Internet Archive - the EatonWeb Portal comprised 957 blogs. At that point, the collection had grown far beyond what any single reader might be able to follow, and the organizational logic of the directory (a non-hierarchical, alphabetical directory of a single Web genre) suggested a sphere, with all points being equidistant from the center, i.e. each blog receiving a link from EatonWeb. As a device, then, the portal served not only to represent the blogosphere and make it navigable, but also shaped it in as an egalitarian space.[3] Meanwhile, Eaton's "inclusive" definition of blogs - for instance, making no distinction between the annotated link lists and personal diaries, as long as these were dated in reverse-chronlogical order - became dominant (Blood, 2000). Following the constructivist argument that in order to exist, the blogosphere had to be actively defined and created, the history of the blogosphere should be seen as

---

[3] The same argument can be made of the Webring, another holdover from the cyberspace period. The adoption of webrings in the blogosphere is discussed in the next section.

tied to that of EatonWeb, and the eventual substitution of the latter by the dominant ordering devices Technorati and Google Blog Search.

By using EatonWeb an historical source set for the early blogosphere, researchers and I chose to amend current approaches to making special collections: the editorial decision of what to include is made retroactively by locating a contemporaneous 'expert list,' and displaced by relying on the portal's inclusive definition of blogs. The obvious drawback to this, however, is that the archiving process is also displaced, and relies on the Wayback Machine. To address the question of archive exhaustiveness, we set out to show the presence and absence of EatonWeb-linked blogs in the Internet Archive. The visualization below was created by taking a screenshot of the first archived version of each page linked to by the EatonWeb Portal in August 2000, and ordering these by date. Sites that were archived as early as 1996 appear in the top left, while those archived in 2000 or later make up the bottom half of the grid. The grey squares in the middle of the grid represent blogs 'missing' from the archive, either because the blog was not indexed by Alexa, or because the site owner blocked the Internet Archive via robots.txt (more on these blogs below). Of the 947 blogs listed by the directory, 857 (or 85.5%) were present in the Internet Archive.

The next step in establishing an historical EatonWeb collection was to specify periods with acceptable sample-sizes and time-frames. For this, we chose to take advantage of a wild-card query in the Wayback Machine that returns the version of a page closest to a particular date.[4] The resulting snapshots capture the homepage of each available blog within a one-year time-frame, returning versions closest to 15 July for each year (1999, 2000 and 2001). Here, the rapid growth of the blogosphere as collected by Eaton is visible: of the 857 archived blogs, just 186 of those were archived in 1999, suggesting that the blogs were not crawled by Alexa in that year or, just as likely, they did not yet exist. (Additional research using historical whois databases may reveal whether the domains were registered at that time.) The query for versions archived in 2000 returned a more representative 764 pages, while that for 2001 returned 775 (see appendix for full list of urls). While this query attempts to concentrate the blogosphere temporally, it still means that up to 12 months might separate the version of one blog from that of another. In other words, these snapshots of the early blogosphere have a long exposure time in order to include a larger sample of blogs. While other collection methods would be used to study specific events (e.g. blogosphere coverage of the September 11 2001 terrorist attacks), the longer range make possible impressions of the general character of the blogosphere over a given period.

How to locate the A-list blogs, and how to characterize or 'profile' the early blogosphere in terms of its associations, or the sources most commonly linked to?

The organization of the EatonWeb Portal suggested egalitarianism, and the theme of inclusivity - a trait attributed to the Web generally, and blogging specifically - was key to many early definitions of blogging (e.g. Blood, 2000; Powazek, 2000). But equally present was a corresponding cynicism regarding media in general (old and new), and the specific media hype around and within the blogosphere. This theme was most clearly demonstrated by counter-hype, or accounts that debunked blogging's promise, especially its supposed egalitarianism:

> The myth, of course, holds that all bloggers are equal, because we all can set out our wares on *the great egalitarian Internet*, where the best ideas bubble to the surface. This free-market theory of information has superficial appeal, but reality is rather different (Clark, 2001; emphasis in original).

Writing in response to a November, 2000 *New Yorker* article on blogging as a growing trend, Joe Clark coined the term 'A-list' to denote "superstar" bloggers receiving attention disproportionate to

---

[4] Wild-card queries are explained in the Wayback Machine FAQ, available at http://www.archive.org/faq.php (accessed 9 November 2009).

their talent for Web criticism (ibid). The image of unremarkable and self-absorbed celebrities no doubt invoked Hollywood and other popular culture, but it is worth noting that the 'A-List' already had a history in cyberculture, used for instance by *Mondo 2000* iconoclast R.U. Sirius to label the Web gurus of *Wired* magazine (Sirius, 1995), and that the problem of elite exclusivity had already been signaled among online diarists, often seen as precursors to webloggers (Eskow, 1997). One difference in the case of blogging, highlighted by Clark, is the significance of the link as a form of recognition and reward. New bloggers would link to the A-List in the hope that it would be reciprocal - since a link from a high-profile blogger meant attracting a large audience - while the A-list bloggers would primarily link among themselves (ibid). Clark goes on to argue that the presence of an A-List, and the resulting linking patterns, contradicts the "nominal purpose" of blogging, namely to provide access to the unfamiliar. In addition to equality among bloggers, the ideal of inclusivity would ensure a variety of sources and points of view (Blood, 2000; Dibbell, 2002). In one sense, the rise of blogging could be seen in a lineage of alternative media, the various independent print and radio initiatives that envisioned a more democratic public sphere. However, the 'independent' label clashes with definitions of blogging as a new form of media criticism, for example linking to a story and providing interpretation or analysis. In this light, one would instead consider the symbiotic (or parasitic) relationship between blogging and news media (Niles, 2007; Carr, 2007).

The two issues - the prominence of the A-list on the one hand and kinds of sources associated with blogs on the other - are equivalent in the sense that they can be measured by the resonance of actors within the blogosphere. Put differently, the (early) blogosphere can be said to be organized around particular dominant actors, whether these be bloggers, news media or otherwise. Using outlinks ripped from each of the three blogosphere snapshots, the clouds below show the most important actors per year, ranked by the number of blogs that link to each. Sites listed on the EatonWeb Portal are highlighted.

(See the Early Blogosphere Actor Clouds, figures 1-3)

A number of things stand out. First, the clouds show the composition of the A-list in each year: while the "celebrity" blogs discussed in the *New Yorker* and elsewhere are present, including robotwisdom.com, kottke.org and megnut.com, the highest-ranked blogs tend to be collaborative (slashdot.org and metafilter.com) or, in 2000 and 2001, act as entry points to the blogosphere (eatonweb.com and jish.nu - the latter was home to the webloggers webring). This is related to a

second finding, and what is perhaps the most striking change, which deals with the rate of standardization in the blogosphere: in stark contrast to1999, when no blog software providers were among the top actors, the highest ranked actor in 2000 and 2001 is blogger.com.[5] Third, the clouds show that in addition to traditional sources of news (the New York Times and the Washington Post), blogs frequently linked to Web and technology-focused news (Wired, Slashdot and ZDnet). If anything, the most prominent sources were more technology-focused in 1999, becoming more general in later years.

How to map the blogosphere, showing clustering by type of blog or around particular sources, and track changes over time?

By showing the resonance of different actors within the blogosphere, context that is unavailable in current single-site and event-based web historiography is made visible. Such a profile puts on display the blogs and other relevant actors that were most widely linked to, demonstrating for instance an acute rise in the importance of standardized blogging software between 1999 and 2000. Taking this focus on the connections between sites further, we created network diagrams for each snapshot, found below.

Where the actor clouds show resonance, the aim of the network analysis is to gauge the relative position of actors within the blogosphere. This is especially useful in the case of sites that are not available via the archive (generally because their owners block the archive from indexing their sites), such as the early A-list blogs camworld.com and bradlands.com. While it is impossible to view versions of their sites from 1999, the analysis makes it possible to estimate their position within the blogosphere at that time. For example, the centrality of camworld.com in the 1999 diagram suggests the high level of influence of Cameron Barrett, who in addition to posting widely circulated essays on blogging in 1998 and 1999 also created the first blogroll, which in turn inspired Brigitte Eaton to start listing blogs on the EatonWeb Portal (Rosenberg, 2009). One major contribution of historical network analysis, then, is to highlight possibly related sites and, to a smaller extent, ‘conjure up’ sites not archived. The same principle is used to identify genres within the early blogosphere: also in the 1999 diagram, tech blogs such as Slashdot and Dave Winer’s scripting.com can be seen to cluster with sources such as CNET and The Standard, news sites focused on the computing industry.

---

[5] However, Dave Winer’s ‘Frontier’ software was available via his site, scripting.com. Frontier can be seen as a predecessor of Web-based blogging software such as Blogger.

(See early blogosphere network diagrams, figures 4-6)

In addition to estimating network centrality, a relational analysis distinguishes between types of actors, showing blogs that link to the network, blogs that link to and receive links from the network, as well as the other actors (unarchived blogs, news media, hosting services, etc.) receiving links from the sample of blogs. Because of this, it is possible to assess to the distribution of links, perhaps point to a more or less egalitarian blogosphere. In the 2000 snapshot, a higher rate of blogs receive links, whereas in 2001 what might be called the "A-List effect" is more clearly visible. Finally, in addition to showing genre, the clustering of actors may also suggest heterogeneity. Whereas for 1999 it is relatively easy to show which blogs are "tech" and which are not, based on their position in the network, in 2000 and 2001 actors such as *Slashdot, Wired* and the *New York Times* cluster together.

Conclusion

The Internet Archive's limitations should be seen less in terms of technical flaws and more as a product of the period in which it was created. Its form and its ambition reveal its embeddedness in a previous period of the Web, cyberculture. By privileging single-site histories and 1990s-style "surfing" through the innovative navigational features of the Wayback Machine, however, the Archive poses challenges for researchers looking to research past states of the Web, as opposed to earlier versions of websites.

The transition from cyberculture to the Web culture of the past 10 years can be traced to the broad rejection of cyberspace as metaphor for experience of being online. One source of this development was the rise of blogging, which offered a vision of the Web inextricably tied up in everyday life, as opposed to it being a space for the extraordinary. Blogging also popularized the notion of the link as reputation-indicator and the assumption that the Web could be 'filtered' in an organic, democratic way.

In researching the early blogosphere, then, we looked to engage with a 'state' of the Web not preserved by the Internet Archive and the Wayback Machine. After locating a historical source set, the EatonWeb directory, and determining the presence and absence of early blogs in the archive, we looked to innovate methods that, like the early blogs, emphasized the importance of the connections between sites, and associations made by linking. In profiling the early blogosphere based on relevant actors, we showed the quickness with which standardized blogging software came to

dominate, and the subtle but important shift in focus toward traditional media, as opposed to Web-friendly news sources. In addition to the medium's massification, then, one thus sees the increasing resonance of mainstream media within the blogosphere. In the network diagrams, we showed how clustering by actors can indicate genre, and suggested ways in which questions of egalitarianism in the early blogosphere might be investigated empirically.

References

Barbrook, Richard, and Andy Cameron. “The Californian Ideology.” *Science as Culture* 26.6 (1996): 1. Print.

Barger, Jorn. “Adding value to your links.” 1999. Web. 4 Nov 2009.

---. “FAQ: Weblog resources.” *Robot Wisdom* 1999. Web. 2 Nov 2009.

Blood, Rebecca. “Introduction.” *We've Got Blog: How Weblogs Are Changing Our Culture*. Ed. John Rodzvilla. Cambridge, MA: Perseus Publishing, 2002. ix-xiii. Print.

---. “Weblogs: A History And Perspective.” *Rebecca's Pocket* 7 Sep 2000. Web. 2 Nov 2009.

Boutin, Paul. “The Internet Archive wants your files..” *Slate Magazine* 7 Apr 2005. Web. 15 Oct 2009.

Brügger, Niels. “Website History and the Website as an Object of Study.” *New Media & Society* 11.1-2 (2009): 115. Print.

Carr, Nicholas. “In praise of the parasitic blogger.” *Rough Type* 5 Mar 2007. Web. 11 Nov 2009.

Clark, Joe. “Deconstructing ‘You’ve Got Blog’.” *Fawny.org* 2001. Web. 10 Nov 2009.

Dibbell, Julian. “Portrait of the Blogger as a Young Man.” *We've Got Blog: How Weblogs Are Changing Our Culture*. Ed. John Rodzvilla. Cambridge, MA: Perseus Publishing, 2002. 69-77. Print.

Eskow, Simon. “Journal-ism.” *Time Digital* 4 Dec 1997. Web. 10 Nov 2009.

Kahle, Brewster. “Brewster Kahle builds a free digital library.” 3 Dec 2007. Web. 16 Oct 2009.

Levy, Steven, and Brad Stone. “The New Wisdom of the Web.” *Newsweek* 3 Apr 2006. Web. 9 Sep 2009.

Lynch, Michael. “Archives in formation: privileged spaces, popular archives and paper trails.” *History of the Human Sciences* 12.2 (1999): 65-87. Print.

Niles, Robert. “Are blogs a 'parasitic' medium?.” *The Online Journalism Review* 2 Mar 2007. Web. 11 Nov 2009.

Powazek, Derek. “What the Hell is a Weblog and Why Won't They Leave Me Alone?.” *Powazek Productions* 2 Feb 2000. Web. 10 Nov 2009.

Rein, Lisa. “Brewster Kahle on the Internet Archive and People's Technology.” *O'Reilly Media* 22 Jan 2004. Web. 13 Oct 2009.

Rhodes, John S. “In the Trenches with a Weblog Pioneer: and interview with the force behind eatonweb, Brigitte F. Eaton.” *WebWord* 29 Nov 1999. Web. 5 Nov 2009.

Rosenberg, Scott. *Say Everything*. Crown Publishing Group, 2009. Print.

Schneider, Steve, Kirsten Foot, and Brewster Kahle. “About.” *September 11 Web Archive*. Web. 13 Oct 2009.

Schneider, Steven, and Kirsten Foot. “The web as an object of study.” *New Media and Society* 6.1 (2004): 114-122. Print.

Schneider, Steven et al. “Building thematic Web collections: Challenges and experiences from the September 11 Web archive and the election 2002 Web archive.” *3rd ECDL Workshop on Web Archives*. Trondheim, Norway, 2003. Web. 31 Oct 2009.

Silver, David. “Introducing Cyberculture.” *Resource Center For Cyberculture Studies* 2000. Web. 24 Aug 2008.

Sirius, R.U. “Mondo 2000 vs. Wired.” *Scrappi.com (accessed via the Internet Archive)* 1995. Web. 21 Jul 2008.

Blogosphere Actors List: a Snapshot from 1999
(Ranked by number of received links. Blogs highlighted.)

Digital Methods Initiative | 21 August | 09

Map generated by tools.digitalmethods.net

wired.com (37) slashdot.org (34)

amazon.com (27) nytimes.com (26) washingtonpost.com (24)

salon.com (23) news.bbc.co.uk (23) cnn.com (23) zdnet.com (22)

sfgate.com (16) abcnews.go.com (16) msnbc.com (16) dailynews.yahoo.com (16)

camworld.com (15) microsoft.com (13) my.netscape.com (13) mercurycenter.com (13) robotwisdom.com (13)

geocities.com (12) news.com (12) salonmagazine.com (12) usatoday.com (11) news.cnet.com (11) newscientist.com (11)

theregister.co.uk (11) members.tripod.com (11) search.washingtonpost.com (10) us.imdb.com (10) latimes.com (10)

members.xoom.com (10) techweb.com (9) boston.com (9) apple.com (8) ntk.net (8) web.pitas.com (8) memepool.com (8) infoworld.com (7) peterme.com (7)

nandotimes.com (7) catless.ncl.ac.uk (7) thestandard.com (7) chicagotribune.com (7) my.userland.com (7) freshmeat.net (7) cgi.ebay.com (7) newsunlimited.co.uk (7)

scripting.com (7) blogger.com (7) forbes.com (7) flutterby.com (6) roosh.com (6) rc3.org (6) eatonweb.com (6) theonion.com (6) cdnow.com (6) google.com (6) observer.com (6)

userfriendly.org (6) byte.com (6) obscurestore.com (6) slate.com (6) herring.com (6) egroups.com (6) villagevoice.com (6) drudgereport.com (5) thestranger.com (5) linkwatcher.com (5) calendarlive.com (5)

mediainfo.com (5) cgi.pathfinder.com (5) nypostonline.com (5) pathfinder.com (5) yahoo.com (5) epinions.com (5) bradlands.com (5) sjmercury.com (5) biz.yahoo.com (5) computerworld.com (5) news.excite.com (5)

pcworld.com (5) thecounter.com (5) suntimes.com (5) mtv.com (5) seattletimes.com (5)

Blogosphere Actor List: a Snapshot from 2000
(Ranked by number of received links. Blogs highlighted.)

Digital Methods Initiative | 20 August | **09**

Map generated by tools.digitalmethods.net

blogger.com (214) salon.com (106)

geocities.com (103) jish.nu (103) nav.webring.org (97) amazon.com (91) cnn.com (83)

dailynews.yahoo.com (82) slashdot.org (79) nytimes.com (72) wired.com (72) eatonweb.com (61)

wrongwaygoback.com (61) news.bbc.co.uk (56) camworld.com (54) kottke.org (54) washingtonpost.com (54)

pitas.com (50) metafilter.com (49) zdnet.com (49) msnbc.com (48) robotwisdom.com (47) news.cnet.com (43) google.com (43)

obscurestore.com (42) theonion.com (41) bradlands.com (40) megnut.com (40) ringsurf.com (39) memepool.com (38) evhead.com (37)

weblogs.com (37) members.tripod.com (36) us.imdb.com (34) scripting.com (34) thecounter.com (34) stormwerks.com (33) webring.org (33) abcnews.go.com (32)

linkwatcher.com (32) theregister.co.uk (32) usatoday.com (32) sfgate.com (31) swallowingtacks.com (31) news.excite.com (30) slate.msn.com (30) harrumph.com (30)

thestandard.com (30) rebeccablood.net (28) latimes.com (27) washingtonpost.com (26) feedmag.com (26) misterpants.com (26) pocketgeek.com (26) zeldman.com (26)

students.washington.edu (25) useit.com (25) web.pitas.com (24) apple.com (24) crosswinds.net (23) egroups.com (23) nandotimes.com (23) powazek.com (23) riothero.com (23)

angelfire.com (22) larkfarm.com (22) wwa.com/~dhartung (22) catless.ncl.ac.uk (21) cafepress.com (21) haughey.com (21) newscientist.com (21) suck.com (21)

tomalak.org (21) espn.go.com (20) sm3.sitemeter.com (20) abcnews.go.com (20) pbs.org (20) peterme.com (20) villagevoice.com (20)

Blogosphere Actors List: a Snapshot from 2001
(Ranked by number of received links. Blogs highlighted.)

Digital Methods Initiative | 21 August | 09

Map generated by tools.digitalmethods.net

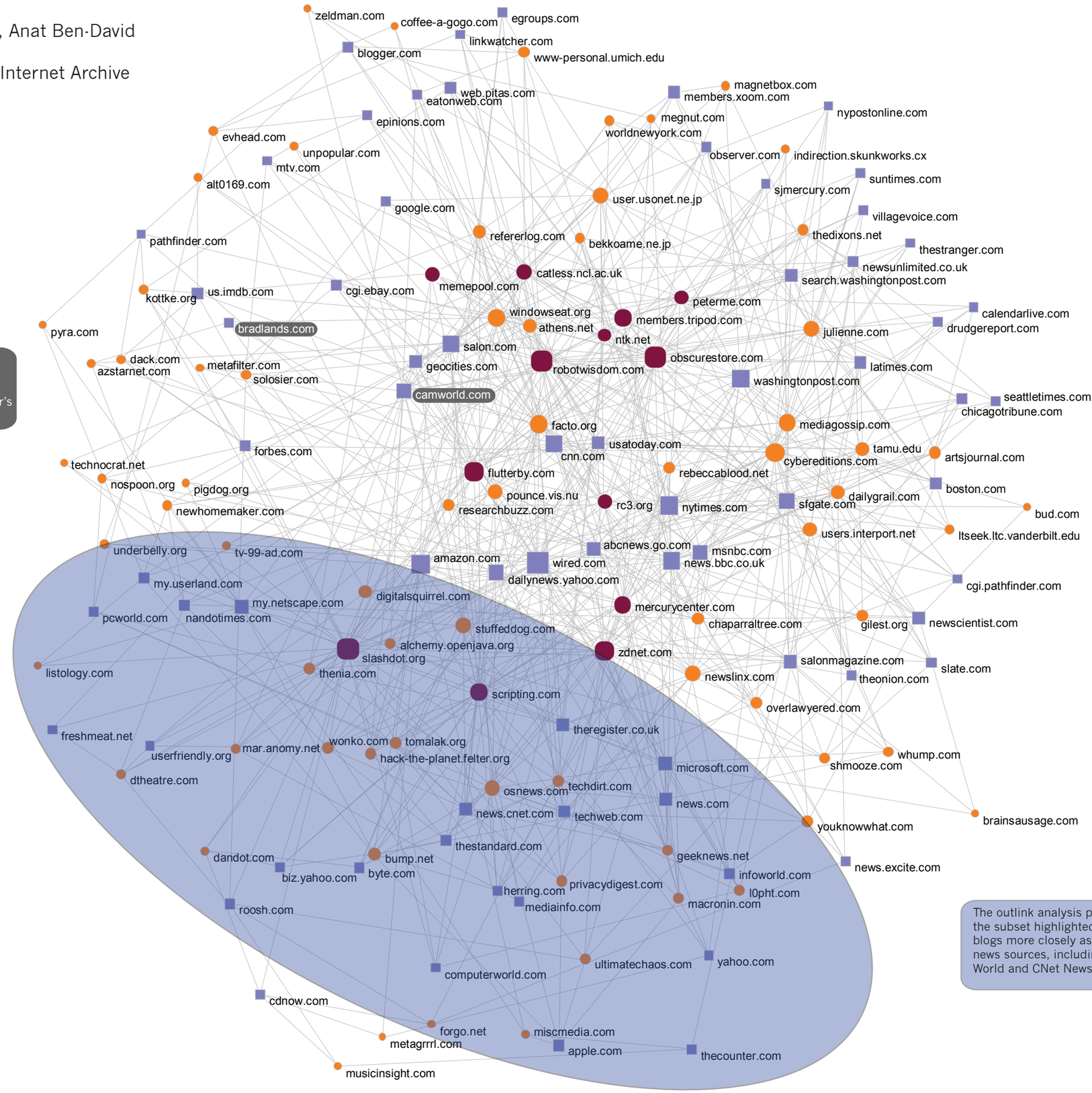
blogger.com (144) amazon.com (107)
dailynews.yahoo.com (97) nytimes.com (87) salon.com (87)
geocities.com (79) washingtonpost.com (73) cnn.com (69) wired.com (69)
news.bbc.co.uk (68) jish.nu (68) slashdot.org (67) google.com (64) metafilter.com (58)
pitas.com (56) msnbc.com (53) ringsurf.com (47) eatonweb.com (43) kottke.org (43) theregister.co.uk (40)
camworld.com (40) scripting.com (38) nav.webring.org (38) latimes.com (37) us.imdb.com (36) robotwisdom.com (36) theonion.com (36)
zdnet.com (35) abcnews.go.com (32) obscurestore.com (32) newscientist.com (32) evhead.com (31) rebeccablood.net (31) zeldman.com (29) guardian.co.uk (28)
usatoday.com (28) noahgrey.com (28) sfgate.com (28) memepool.com (28) boston.com (27) members.tripod.com (27) slate.msn.com (27) news.cnet.com (26) webstandards.org (25)
cafepress.com (25) groups.yahoo.com (25) time.com (24) wrongwaygoback.com (24) peterme.com (24) ananova.com (24) linkwatcher.com (23) thecounter.com (23) theatlantic.com (23) csmonitor.com (23)
misterpants.com (22) megnut.com (22) apple.com (22) weblogs.com (22) larkfarm.com (22) v1.nedstatbasic.net (21) villagevoice.com (21) chicagotribune.com (21) harrumph.com (21) groups.google.com (20)
members.aol.com (20)

The Archived Blogosphere: a Snapshot from 1999

Method_ Gather outlinks from EatonWeb blogs archived in 1999, using the version closest to July 15. Perform cluster analysis and visualize network.

Analysis_ Esther Weltevrede, Carolin Gerlitz, Anat Ben-David and Michael Stevenson
DMI Summer '09_ Digital Methods and the Internet Archive

The outlinks analyzed are directed at both archived and non-archived blogs, making it possible to estimate the latter's position within the network.

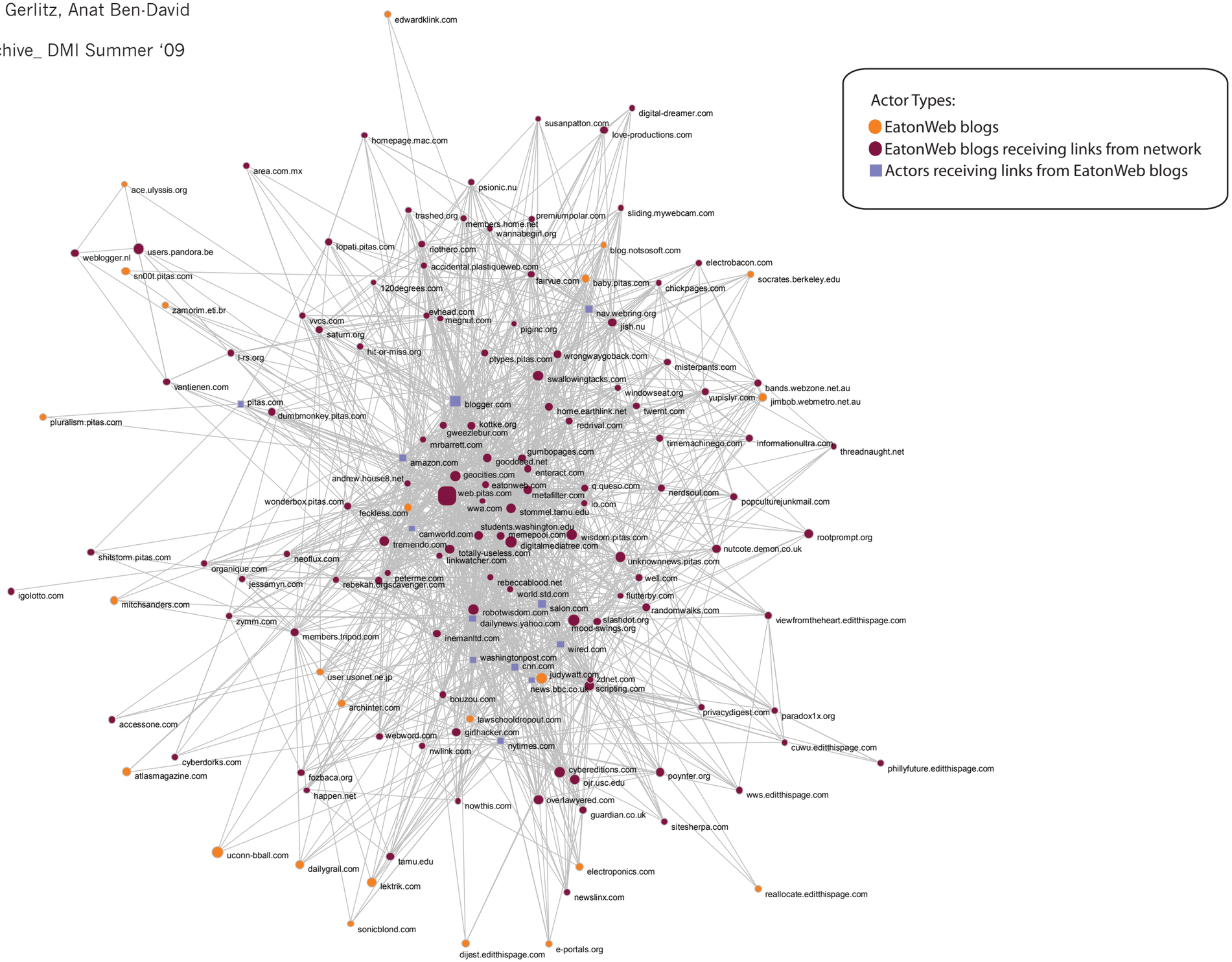


The outlink analysis produces clusters: the subset highlighted here includes blogs more closely associated with tech news sources, including Computer World and CNet News.

The Archived Blogosphere_ a Snapshot from 2000

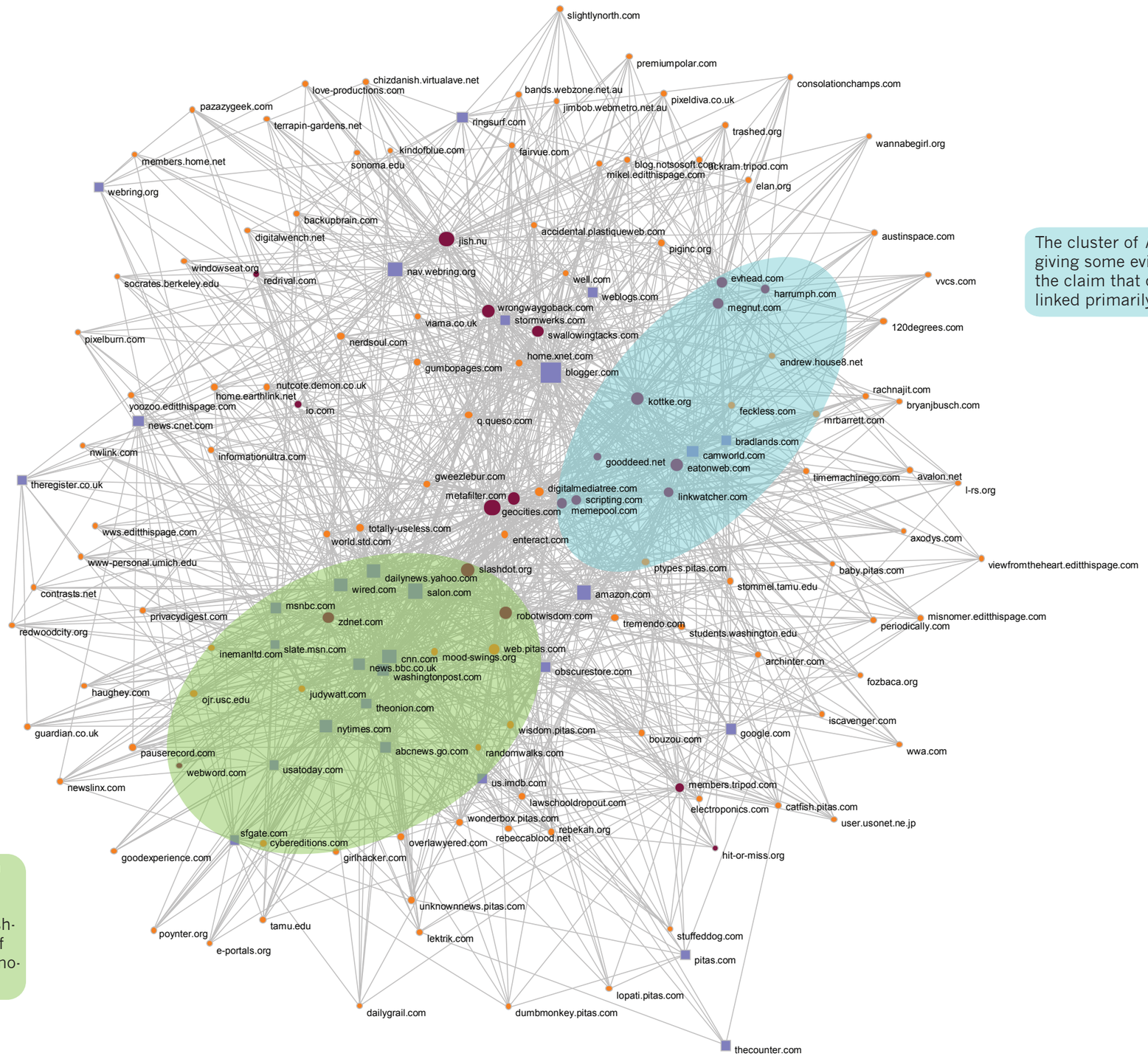
Method_ Gather outlinks from Eaton Web blogs archived in 2000, using the version closest to July 15. Perform cluster analysis and visualize network.

Analysis_ Esther Weltevrede, Carolin Gerlitz, Anat Ben-David and Michael Stevenson
Digital Methods and the Internet Archive_ DMI Summer '09



The Archived Blogosphere_ a Snapshot from 2001

Method_ Gather outlinks from EatonWeb blogs archived in 2001, using the version closest to July 15. Perform cluster analysis and visualize network.
Analysis_ Esther Weltevrede, Carolin Gerlitz, Anat Ben-David and Michael Stevenson
Digital Methods and the Internet Archive_ DMI Summer '09



The cluster of A-List blogs is visible, giving some evidence in support of the claim that celebrity bloggers linked primarily to one another.

In 2001, the sources for news in the blogosphere were no longer predominantly 'Web' or tech-focused, such as Wired and Slash-dot. Instead, there is a cluster of traditional, Web-based and techno-logical news sources.