

Towards a hyperlinked society: a critical review of link studies.

Abstract: The hyperlink is a fundamental feature of the web. This paper investigates how hyperlinks have been used as research objects in social sciences. Reviewing a large body of literature belonging to sociology, political sciences, information sciences, geography or media studies, it particularly reflects on the study of hyperlinks as indicators of other social phenomena. Why are links counted and hyperlink networks measured? How are links interpreted? The paper then focuses on barriers and limitations to the study of links. It addresses the issue of unobtrusiveness, the importance of interpreting links in context, and the possibilities of large-scale, automatic link studies. We finally argue that the studies reviewed make a powerful case for mixing methods. Link studies are a supplementary tool in the social science researcher's toolbox: they should be used when relevant, and are only worthwhile when triangulated with other web-based and non web-based methods.

Keywords: hyperlinks, hypertextuality, hyperlink network analysis, link analysis, web studies

Hypertextuality has always been a fundamental characteristic of the world wide web since its inception in the early 1990's. The ability to link pages, sites and documents stands out as what essentially differentiates the web from other media (Napoli, 2008: 56). For many commentators, hypertext constitutes a major shift. However, there is more to hypertextuality than its mere technical aspects: it affects communication, communities and the public sphere at large. "Hyperlinking has changed the fundamental dynamics of human communication" declares Hespos (2008: 137), while Turow maintains that "linking affects the overall size and shape

of the public sphere” (Turow, 2008: 4). Others claim that hyperlinks are highly loaded with symbolic and social power:

“Hyperlinks are the glue of these online communities, forming digital footprints of the way individuals make connections. Through a simple selection to include, exclude or just follow a link in our daily online interactions, we passively telegraph the way we see the world, what is important to us, to what degree and why” (Schulman, 2008: 147).

In short, hyperlinks are more than technical artifacts as we increasingly live in a “hyperlinked society” (Turow and Tsui, 2008).

This paper shines a light on hyperlinks as a research objects, and reviews a vast array of studies that have examined links. It will especially concentrate on inquiries into hypertextuality as a proxy for another social phenomenon, and will reflect on why and how links are studied. As we live in the so-called *hyperlinked society*, it is no surprise that many researchers have spent considerable time and energy studying hyperlinks. Research scrutinizing links expands in various fields and domains: political science, geography, economics, media studies, or sociology. But what exactly are we trying to grasp when examining links? When and how is the study of links relevant to answer a research question?

Studying the link

First, we must distinguish two streams of research: the literal and the interpretational. The former aims at describing hyperlinks *per se*; at discovering

their properties, while the latter uses links as indicators of other phenomena. The literal trend investigates hypertextuality for its own sake, in order to understand how links are structured and organized. For instance, studies pertaining to the “new science of networks” (Barabasi, 2003; Watts, 2004a; Watts, 2004b) throw light on properties of hyperlink networks at a general level. They are merely interested in understanding the web’s underlying structure. Such investigations into global hyperlinking patterns have “uncovered principles that help to understand networks of all type” (Tremayne, 2004: 234). “Peculiar and fascinating properties” (Ghitalla, 2009) were identified. For example, researchers have determined that the web is a “scale-free network, dominated by hubs and nodes with a very large number of links” (Barabasi, 2003: 165), and have revealed “strong regularities, among which the existence of a ‘universal power law’” (Adamic and Huberman, 2001: 131). A power law distribution is “a distribution for which there exists a scale-free inverse relationship between an observed magnitude of an event and its frequency” (Halavais, 2009: 198). This topology goes around with Pareto’s law, or the so-called 80/20 rule (Tremayne, 2004: 238): 20% of the nodes possess 80% of the links – a small part of websites collects the majority of links while most receive few or no links. For their part, Albert et al. (1999) established the diameter of the web and its small-world nature: despite the impressive total amount of existing websites, there are, on average, nineteen

degrees of separation between two pages (Barabasi, 2003: 165), i.e. one webpage is only nineteen clicks away from another.

Beside these results emphasizing low distance as a characteristic of the web, researchers demonstrated that finding a path between two pages is not that obvious: Broder et al. (2000) proposed a fragmented model of the web corresponding to a bow-tie structure where the direction of the links strongly matters (Barabasi, 2003: 166). According to their analysis, the web's macroscopic structure naturally breaks into four parts: there is a central core, coined the "giant strongly connected component", all of whose pages can reach one another; an IN-component that can reach the core; an OUT-component that can be reached from it; and tendrils that can neither reach the core nor be reached from it (Broder et al., 2000). This set of findings belonging to the science of networks can be compared to physical laws: they form a collection of general observations that one needs to be aware of when dealing with hyperlinked material.

The social significance of links

All the above are inquiries into hyperlinking as an interesting phenomenon *per se*. But the majority of investigations into links go further: links are often used as substitutes for something else, as indicators of other phenomena such as authority, trust between parties involved in commercial transactions, political

affiliations, campaign strategies, power relations between countries, social networks, the quality of the research done by universities, transparency in the practice of journalism, etc. Those studies perpetuate the idea that linking behavior is not random (Rogers and Marres, 2000: 6), and that links are “socially significant in some way” (Hsu and Park, 2010). In this perspective, links have an “information side-effect”, they can be used to understand other facts even though they were not individually designed to do so: “information side-effects are by-products of data intended for one use which can be mined in order to understand some tangential, and possibly larger scale, phenomena” (Adamic and Adar, 2003). In a word, there is a “widespread belief that hyperlinks between web pages can yield useful information” (Thelwall, 2006: 1). Two underlying assumptions sustain this claim, at an individual and global level. On the one hand, it means that every link has been created for a precise purpose, that webmasters link selectively (as opposed to capriciously) in accordance with a “politics of association” (Rogers and Ben-David, 2008: 43): “Making a link to another site, not making a link, or removing a link, may be viewed as acts of association, non-association or disassociation, respectively” (Rogers, 2010). On the other hand, links make sense at a global, aggregated level: taken together, they reflect social and cultural structures (Halavais, 2008: 39). In other words, “a hyperlink is not only a link but has certain sociological meanings” (Hsu and Park, 2010: 2).

Indicators of authority

For what other phenomena do hyperlinks stand? Which sociological meaning hides behind them? The work of Brin and Page (1998) in which they explained the logics of their “Page Rank” algorithm initiated the trend of using links to calculate authority or relevance (Rogers, 2010). In this perspective, every link is considered as an implicit vote about the quality or relevance of a piece of content (Finkelstein, 2008): the more a piece of content is linked, the more authoritative it is. Such assumption rests on a parallel between hyperlinks and citations in the academic context: when often cited by their peers, scientists’ contributions are considered important and valuable. Similarly, Brunn and Dodge (2001) argue that we can exploit hyperlinks to assess trust in the context of international online trade. In concrete terms, various methods and different indexes have been proposed to systematically measure this notion of authority or trust on the web, as illustrated by the famous Page Rank (Brin and Page, 1998) used by Google or the “authorities/hubs” model (Borodin et al., 2001). Both rely on the idea that often-linked nodes possess some level of authority that is subsequently redistributed on other nodes. Other measurements directly draw on the study of social networks (Scott, 2000), where it is common to assess a node’s centrality (its importance in terms of connections) through various indicators;

even if the notion of authority is less pregnant, the centrality of a node denotes its prominence in the network.

Monitoring academic performance

Research also expands on the parallel between links and citations, and broadly uses links to monitor various aspects of academic performance. Information sciences have extensively explored link analysis and applied its methods, tools and techniques to academic links (Thelwall, 2004). In this respect, Thelwall (2001) built a link-based “web impact factor” to evaluate the research activity of universities in Britain. Harries et al. (2004) mapped the bounds between academic disciplines, with the underlying goal of evaluating universities and identifying patterns in the conduct of research. As for Park (2010), he investigated the “semantic variation, disciplinary scope and institutional structure of e-science programs in South Korea” (Park, 2010: 211), using links as proxies for relationships between institutions and stakeholders (Park, 2010: 223). Those authors, however, remain wary of such “webometrics” approaches. As Thelwall (2001) underlined, web link metrics can yield results that highly correlate with non-electronic phenomena (such as traditional impact factors in academe) but as with traditional citation analysis, it will never fully replace traditional measurements: those who only rely on link counts to judge academic performance

risk serious errors (Thelwall, 2001: 18). Moreover, he insists on the need to combine link studies with non web-based methods, in order to ensure a truthful evaluation of universities.

Reflecting political affiliations

Link studies have also gained importance in the field of political science, where researchers have investigated the “political web”. What do links mean in this context? Papers by Park and other colleagues continuously explored South-Korean politicians’ web presence. They mined links to assess the “new ecology of political communication” (Park et al., 2005). Their main idea is that links signal political affiliations, whether they investigate links stemming from politicians’ homepages to understand their political agenda, the communication network formed by assembly members and political parties (Park et al., 2004), or the relations between citizens’ and politicians’ blogs (Park and Jankowski, 2008). For instance, “links embedded in political blogs can be taken as technical indicator of the ideological landscape of a blog sphere being studied” (Park and Jankowski, 2008: 62). Still in the political field, Williams et al. (2005) examined the hyperlinking strategies of candidates taking part in the 2004 US presidential race, and compared the different strategies used in the candidates’ blogs or homepages. They suggest that different genres of websites go with different linking styles: links

on the candidates' homepages more often point to internal fundraising or revenue-generating pages, whereas blog links are less likely to solicit donations and propose more external links. In this case, the authors suggest, links are loaded with strategic campaign interests.

Signs of political homophily

When it comes to assessing political issues, link studies are additionally concerned with the notions of “political homophily” and “cyberbalkanisation”. Proposed by Sunstein (2001), the notion of cyberbalkanisation suggests that with the possibilities of personalization proposed by the web and its ever-growing offer in contents, people tend to generally consult pages harboring opinions similar to theirs. In Sunstein’s view, such isolation is a danger for democracy, as “people will abandon the reading of dissenting political opinions in favor of material that is closely aligned with their own ideological position” (Hargittai et al., 2008). Hargittai et al. (2008) tested the cyberbalkanisation assumption in the 2004 US political blogosphere, and found that widely read political bloggers link more often to those who share their views (Hargittai et al., 2008). But this trend, they added, does not increase over time. Also in the context of the 2004 U.S. Presidential campaign, Adamic and Glance measured the interconnectedness of the two parties, as embodied by the “degree of interaction between liberal and conservative blogs”

(Adamic and Glance, 2005: 1). They witnessed a “divided blogosphere: liberals and conservatives linking primarily within their separate communities” (Adamic and Glance, 2005: 14), with only 10% of the links bridging the two groups (Adamic, 2008: 230). Moreover, they detected that conservative blogs link to each other more frequently and in denser patterns. Ackland et al. (2009), further exploring Adamic and Glance’s dataset, claimed that manifestations of homophily – or “assortative mixing” as they call it – must not be attributed to the intrinsic nature of bloggers to favor their own kind, but that this linking behavior is explained by a model of network formation in the blogosphere and, more precisely, by the dynamics between a political majority and minority. In other words, homophily is merely an effect of the forces of the network, and “it is possible to view the creation of links by and between political bloggers as a reflection of the relative popularity of one ideology over another in the underlying population from which bloggers are drawn” (Ackland and Shorish, 2009). The authors admitted, however, that while focusing on link creation as a result of the relative popularity of one ideology over another, their model abstracts away from other influences likely to dictate link formation (Ackland and Shorish, 2009: 397). The evolution of political linking patterns over time has also been studied, as exemplified by Hsu and Park who scrutinized the differences between different generations of links. Their “sociology of hyperlinks” explored links produced by South-Korean politicians, and

they highlighted a significant evolution between web 1.0 (personal websites), web 2.0 (blogs) and Twitter links: the network of links appeared denser over time, and there is evidence that it became increasingly less centralized (Hsu and Park, 2010: 10). “Prominent Web 1.0 hubs with many links tended to disappear, but butterfly networks based on political homophily emerged” (Hsu and Park, 2010: 1). To sum up the state of link studies in the political field, we see that researchers use links as indicators of other phenomena that vary in their levels of abstraction and generalization. The most down-to-earth approach interprets links as mere signs of communication networks between involved actors, or strategic tools for campaigning in one particular case study. One step further, we find the notions of links as signs of ideological affiliations, or symptoms of “cyberbalkanisation”.

Tracing public debates

Besides politicians’ online behaviors, wider political issues are also on the agenda. Here, scanning links involves looking at how public debate is shaped on the web, and how “discursive affinities” or “issue networks” are structured. Rogers and Ben-David (2008) examined the issue network around the Israeli security fence debate, and showed how the various actors were situated toward each other. By mapping the links between NGOs’ websites, they discovered that Israeli left-wing organizations are in “virtual isolation” in the overall issue space (Rogers and

Ben-David, 2008: 63). Similarly, Rogers and Marres (2000) attempted at “debate-scaping” the discussion about climate-change. They encounter different “linking styles” typical of the .com, .gov and .org domains, and noticed, for example, that “governmental institutions tend to form one massive, autonomously operating body” (Rogers and Marres, 2000: 9). Consequently, they argue that the debate about global climate change is owing to the presence of a central governmental authority (Rogers and Marres, 2000: 13). Marres (2004), in turn, concentrated on the argument around a controversial project led by the World Bank. Elaborating on actor-network-theory, she used links to describe how the issue was evolving and shaped by different actors. All those studies use the links as visible trails of public issues being debated, as tools to clearly map the positions of involved actors. The authors implicitly suggest that such debate-scaping is useful to better understand what is at stake in our democratic societies.

Connections between blogs

Efforts to map links between blogs also relate to this movement. In this context, hypertextual proximity helps to make sense of vast “blogospheres”, i.e. sets of interconnected blogs. Blogs are particularly salient objects when it comes to studying links, because they are a link-driven genre (Blood, 2000). Bruns (2007) mixed the issue-centered approach with the emphasis on blogs, as he maps the

interactions between Australian bloggers about the David Hicks case. His findings highlighted a “tendency for discussion to cluster around a handful of sites which are defined by their political orientation”, a lack of coverage of the case by bloggers outside of Australia as well as limited interactions between bloggers and the mainstream media (Bruns, 2007). Tremayne et al. (2006) mapped the structure of the Iraq war blogosphere. Their findings are consistent with previous comments on cyberbalkanisation, as they highlight a blogosphere structured in two distinct halves, the liberal and the conservative – “but also a fairly robust cluster of blogs that serve as conduits between the sides” (Tremayne et al., 2006: 305). Etling and his colleagues’ approach is wider: they scanned nation-wide blogospheres at a global level (Etling et al., 2009; Etling et al., 2010; Kelly and Etling, 2008) to understand how public debates between bloggers are shaped and how they fit in a wider media ecology. In that perspective, links significantly help in detecting “attentive clusters”, i.e. groups of blogs that show similar linking behavior and that link to the same resources. Those methods are “based on the principle that macro structure arises from the tendency of individuals to link more frequently to things that interest them and to people with whom they share attributes or social relations” (Etling et al., 2010: 11). Researchers use linking patterns to make sense of large amounts of blogs and to gather them in significant clusters – as similar linking behaviors “often arise from intuitively grasped interests, ideologies,

preferences and affiliations that exist among bloggers” (Etling et al., 2010: 14). From then on, hypertextuality constitutes a non-semantic tool to discover semantic patterns (Jacomy and Ghitalla, 2007: 4): if two blogs link to exactly the same resources, it is likely that they will cover similar topics, and perhaps hold similar views on them. In substance, hyperlinks help to detect “topical localities” (Ghitalla, 2009); hypertextual proximity indicates similarities in content (Ghitalla et al., 2005: 8).

International flows of information

Links studies also set out to tackle journalistic issues. In this framework, hypertextuality is associated with positive journalistic values such as interactivity, transparency, credibility or diversity. Briefly, adding links to stories allegedly improve online news:

“Choosing links to include in your story gets to the very essence of what it means to be a journalist. You are searching out information, assessing and contextualizing it and then presenting it to the user in an understandable, compelling way. Journalists, through the skillful searching, assessing, contextualizing and presenting of linked information, can be modern-day trailblazers, contributing in small, individual ways to making sense of the mass of information available online” (Foust, 2009: 161).

Therefore, links stemming from news sites are investigated to assess those sites’ compliance with the shared professional standards of quality online journalism. Nevertheless, empirical findings tend to indicate that news sites do not

live up to the expectations, as their use of external linking is scarce (Dimitrova et al., 2003; Dimitrova and Neznanski, 2006; Kenney et al., 2000; Quandt, 2008; Stray, 2010; Tankard and Ban, 1998; Tremayne, 2005; Tsui, 2008).

Links on news sites are also examined to map international information flows. Chang et al. (2009) tested the structure of links originating from news sites in regard to the world-system theory “which emphasizes the asymmetry between information-rich and -poor countries” (Barnett et al., 2010: 2). They note that contrary to the impression of openness favored by the internet as a borderless medium, linkages between nations via outgoing hyperlinks to websites of other countries remain relatively closed (Chang et al., 2009: 155). Halavais (2000) presented similar findings concerning the web at large. He argues that most links are restrained within national borders: “although geographic borders may be removed from cyberspace, the social structures found in the ‘real’ world are inscribed in online networks” (Halavais, 2000: 7). The intertwining of “real world” geography and international link structures stands at the core of other studies. Barnett and Sung (2005) examined the patterns of communication among nations, through the interconnection of top-level domain names (such as “.be” or “.fr”). They show that hyperlinks comply with a center-to-periphery structure, in accordance with world system theory. According to them, economic relations are

the primary organizing mechanism of international communication. Links reflect power interactions between countries, and core countries are clearly central to the web. Typically, the United States constitute the most central node in the hyperlink network (Barnett and Sung, 2005: 226). This phenomenon turned out to be even clearer when Barnett et al. (2010) overcame an inherent bias of previous studies by unveiling what is behind the .com domains: the core countries' domination becomes stronger. Also mixing virtual and physical geography, Lin et al. (2007) analyzed links between blogs as indicators of connections between U.S. cities. They pinpointed "super metropolitan clusters" (Lin et al., 2007: 21) i.e. densest hyperlink networks around cities with cultural and political prominence. Thus, "centers of cultural and news production still attract the most attention nationwide" (Lin et al., 2007: 22), whereas cities of less cultural-political significance mostly connect with places near them.

Finally, links are also used as signs of relationships between people. In Adamic and Adar's work (2003), for instance, hyperlinks between homepages are used as an indicator of "friendship" between students. The authors argue that using links in this context allows collecting data to track social network in a convenient way – whereas traditional social network research has to cope with long interviews and observations.

Barriers and limitations to link interpretation

When considering the scope of research projects outlined above, one may think that link studies constitute the panacea for answering all kinds of research questions. But important limitations exist and need to be mentioned. The first promise of link studies relies in their apparent technical simplicity. In theory, links are easy to detect: they are well-defined strings of HTML code. As the HTML code for every webpage is publicly available, there is no obvious barrier to the collection of links. One only has to browse through the code and to retrieve everything between the `<a href>` and `` tags. Doing this at a large scale, however, is time-consuming: therefore, tools exist to facilitate such tasks. “Web-crawlers”, for example, are designed to follow and index links automatically. Pointing them all out would be impossible, but a handful of freely available tools oriented towards social science research and integrating link analysis in their features exist: SocSciBot¹, LexiURL², the VOSON project³, IssueCrawler⁴, Navicrawler⁵, etc. The impression that hyperlinks are out there, ready to be collected and analyzed should not, however, lead us to the assumption that studying links is simple and straightforward.

A false feeling of unobtrusiveness

When discussing the strength of link analysis, Park (2003a) argues that its main quality is that it is unobtrusive and that “hyperlink data can be gathered

naturally without intruding in the research context". However, Halavais stresses that "the trick with human systems is that they evolve and adapt not only to their environment but also to the observer who study them. (...) those measuring the system have become a significant part of it" (Halavais, 2008: 48). Therefore, measuring hyperlink network is not as unobtrusive as it seems (Halavais, 2008: 50). The mere fact that it is widely known that Google's search engine algorithm draws on links creates a potential bias. Search engine optimization is more and more common: "the study of macrolevel hyperlink networks has been complicated by users becoming more aware of hyperlink structures" (Halavais, 2008: 53). Ghitalla also remarks that his hyperlink map of the European web, when published online, logically appears at the center of the network pictured – as it contains links to every other concerned website: "the map, once published on the network, modifies the territory it represents" (Ghitalla, 2008). But there are other ways in which the observer, who thought he was invisible, influences what he observes. In Marres' case (2004), the crawl for links did not go unnoticed and resulted in an e-mail from the owner of the website she was interested in. This reaction, the author argued, "could lead us to pose all sorts of questions, for example about the inevitable involvement of Internet researchers in the phenomenon studied (the anthropologist's problem of 'meshing with the natives')" (Marres, 2004: 125).

Context matters

Another important limit could be stated as follows: all links are not created equal. All the works presented above somehow deal with a paradox: they attempt to make sense of links while recognizing that interpreting them is highly sensitive and difficult. Link interpretation often implies to infer something about the intentions of the people that have created them. Even when considering the simplest dichotomy it is obvious that a link can equally mean praise or criticism. There is no guarantee that indication of the author's intentions can be found in the link itself, or in the linked resource: "it is fundamentally impossible to determine the reason for the link based upon the target page only" (Thelwall, 2001). Zimmer (2009) compares links with *renvois*, an 18th century "system of cross-references" featured prominently in Diderot and d'Alembert's *Encyclopédie* (Zimmer, 2009: 96). He states that these antecedents of hyperlink text are not as straightforward as they seem: the writers of the *Encyclopédie* employed *renvois* to avoid censorship when discussing controversial topics, relying on "irony, innuendo and indirection" (Zimmer, 2009: 103). Similar practices exist in the context of linking. For example, how to make sense of the ironic linking practice known as "Rickrolling"? This famous internet prank happens when a person provides a hyperlink that they claim is relevant to the topic at hand, but the link actually takes the user to a music video for Rick Astley's song *Never Gonna Give You Up* (Wikipedia, n.d.). The nature

of such link is equivocal and we can only understand it if we are aware of the *Rickrolling* meme.

In a word, “the universal nature of hyperlinking makes it a very difficult sort of artifact to understand. The question of what someone means when they create a hyperlink or when they activate one is entirely determined by the context of the hyperlink’s use” (Halavais, 2008). Most authors thus acknowledge the importance of the context of links, explicitly or implicitly. For instance, scholars studying blog links often make a difference between “in-post” links and “blogroll” links, i.e. the links present within the text of a blog post or those displayed in a side-column (Adamic, 2008; Lin et al., 2007), and argue that they cannot be amalgamated. Which links should be favored, and how the different types of blog links should be interpreted? The answer depends of the type of research carried out: those interested in studying affiliation tend to value blogroll links, whereas research into conversation networks concentrates on in-post links. Adamic (2008) argues that blogroll links serve as badges that bloggers may display, they indicate “a general social awareness on behalf of the author” (Marlow, 2004: 3), and constitute an indicator of the ideological landscape surrounding the blogs (Park and Jankowski, 2008; Park and Thelwall, 2003). For Lin et al., blogroll links are more indicative of interpersonal affiliation than in-post links (Lin et al., 2007: 17), whereas in-post

links are considered as a better indicator of active conversations going on in the blogosphere. They allow “authors to have a sort of distributed conversation” (Marlow, 2004: 3), and they “signal active engagement with another blogger” (Hargittai et al., 2008). As argued by Adamic and Glance (2005: 7), the “linking behavior within posts is more indicative of a blogger’s reading activity than are blogroll links”. To sum up, there is no unified theory about the potentially different meanings of blogroll and in-post links.

More broadly, differences between web genres also matter, and the universal nature of link analyses is questioned. For instance, the framework used to make sense of links on personal homepages is deemed inadequate for hyperlinks in the blogosphere “because there are several types that are unique to blogs” (Park and Thelwall, 2008). Moreover, in their attempt at mapping links between different academic subjects, Harries et al. notice disciplinary variances in the types of links, and again underline that the context is crucial. The need to understand the motivations behind linking appears as a leitmotiv. The parallel with academic citations is again prominent, with Thelwall (2003) elaborating a typology of the motivations to link on the basis of previous research about motivations to cite. In their study of the evolution of linking behavior, Hsui and Park noticed that incentives to link vary: a link on Twitter, for instance, is likely to

result of a different motivation than a link on a homepage or on a blog. Most authors acknowledge that grasping the motivations behind linking would greatly enhance their analysis (Park, 2010): “several methods need to be employed to examine the reasons developers of websites form a network with other sites via hyperlinks: survey, in-depth interviews, observation, comparative analysis of website contents and other network data would contribute to an understanding of the social relationships among the network’s components” (Park, 2003a: 58). But by interviewing link creators, notes Thelwall (2006: 4), it would be almost impossible to systematically identify what someone has in mind when placing a link. Concretely, he adds, finding the author of a webpage might be difficult; the authors may not remember why they included URLs; and in-depth interviews are difficult to carry out on a large scale. Besides, even if our understanding of linking motivations were perfect, it would not tell us anything about how the users react to the links. The sheer existence of a link does not even guarantee that the users indeed follow them (Bruns, 2007), or does not ensure that their interpretation of that link is unequivocal (Hargittai, 2008). To put it briefly, one should be very careful when interpreting the social significance of links.

Promises of automation

In this respect, one of the main attractive aspects of link studies should be considered with caution: the possibility to crawl the web at a large scale, to harvest a large amount of links and to study them globally. If all links are not created equal, the automatic treatment and interpretation of a large amount of links may therefore become delicate. In their attempt at studying the whole “web graph”, Bharat et al. (2001) used the link structure to determine if sites are related. But something as simple as link frequency to determine relatedness is sometimes misleading, as the authors witnessed “strongly connected host pairs that didn’t seem to be otherwise related” – such as large hosts like www.geocities.com and members.aol.com that are strongly connected only by virtue of their immense size (Bharat et al., 2001: 5). Thelwall summarizes the issue as follows: “the reason for high link count can be completely different to the average reason for link creation” (Thelwall, 2006: 9). More radically, he pinpoints an important issue regarding the use of statistic: “fundamental statistical problems can arise if links are not the primary object of study but serve as a vehicle to draw conclusions about processes underlying link creation, which is typical of social sciences link research” (Thelwall, 2006: 7). Links are not statistically independent observations, “because of factors such as imitation between web authors (...), the copying of pages or parts of pages, and automatic creation of web pages by server software” (Thelwall, 2006: 7).

According to Thelwall, this means that inferential statistics, such as confidence intervals, are not valid. As Harries et al. (2004) phrase it, web link creation is a “social activity that inspires imitation, the opposite of statistical independence” (Harries et al., 2004: 439). Even without completely dismissing link statistics, we should always remember that most statistical tests assume that the dataset complies with a normal distribution, whereas link distributions globally obey power laws (Barabasi, 2003).

Conclusion: mixing methods

Facing these difficulties in treating links automatically and at a large scale, most researchers opt for semi-automatic methods, or a mix between large-scale, automatic data retrieval and smaller, manual processing. For example, Etling et al. (2010) mainly drew on automatic methods in order to map the “attentive clusters” of the Russian blogosphere – that is, the categories they used to classify blogs relied solely on the linking patterns and were therefore automatically determined – but they combined this method with a manual, qualitative assessment of 1200 blogs. Most of the projects presented above rely, at some point, on manual coding, qualitative appreciations and human expertise. Hargittai et al. (2008) qualitatively qualified links between liberal and conservative blogs; Park (2010) manually classified sites by authors; Park and Jankowski (2008) qualitatively analyzed blog

titles ; Thelwall (2003) manually categorized links by motivation. Some authors especially insist on the role played by experts: only people that are specialists in the field investigated can make sense of the links (Ghitalla, 2008; Jacomy and Ghitalla, 2007).

Most of the papers reviewed here do more than pure link counts or link network analyses; they often enhance quantitative network analyses with qualitative tools aimed at making sense of the links. Moreover, they often mix link counts with other methods and other datasets. Content analysis of texts is often brought in: Adamic and Adar (2003) looked at the texts on presidential candidates' homepages. Etling et al. (2010) use computer-assisted text analysis of blog contents. TACT (*Textual Analysis Computing Tools*) were also used by Rogers and Marres (2000) in order to locate key phrases in their context. Park and Jankowski (2008) carried out a follow-up qualitative analysis on blog titles. Rogers and Ben-David (2008) chose to pair their link analysis with a term analysis: they scrutinized the words used to frame their topic of interest (i.e. the building of the Israeli "security fence"). All those approaches mix link studies with the examination of other web contents, and other methods. For Thelwall (2001), however, this is not enough: we need not only to combine link study with the analysis of other web contents, but also to combine it with "non-web based information", in order to go

“towards a hybrid calculation combining web information with another source”. For example, in his study of universities’ “web impact factor”, Thelwall integrated the number of Faculty members of a particular institution (i.e. a non-web based information) as a factor in the calculation.

The appeal of link studies mainly relies on their simplicity: links are numerous, easy to identify and to collect, as well as commonly used on the web. Through a review of a vast array of research projects involved in link analyses, we have nevertheless shown that link interpretation is convoluted and sometimes problematic. However, the variety of fields applying link studies, as well as the multiplicity of possible interpretations should not lead us to conclude that these tools and methods are chaotic or sterile. Several pragmatic conclusions must nevertheless be pointed out: a univocal, unambiguous interpretation of the link does not exist. Similarly, the possibility of automatically handling large amount of data should be limited to certain areas, such as the science of networks – that merely aims at determining how the network of links is organized. When starting a research project involving link analysis, one should always ponder the following question: “Where does social research end, and Google studies begin?” (Rogers et al., 2009), i.e. when do we stop to purely describe the structure of links as technical objects and when do we start to relevantly exploit links to make sense of a social

phenomenon? If we choose the second option, we must ensure validity by combining link analysis with other tools and methods, as well as with a strong focus on human expertise to make sense of the connections. Since the advent of the hyperlinked society, connections between actors embodied in websites and webpages have been explicitly visible and we should not let such opportunity unstudied. Tracing the links, however, is not an end *per se*. It should rather be considered as what it plainly is: a supplementary tool in the social researcher web-oriented toolbox – useless on its own, but extremely powerful when manipulated competently.

References

- Ackland R and Shorish J (2009) Network Formation in the Political Blogosphere. An Application of Agent Based Simulation and e-Research Tools. *Computational Economics* 34(4): 383-398.
- Adamic L and Glance N (2005) The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *LinkKDD '05: Proceedings of the 3rd international workshop on link discovery*: 36-43.
- Adamic LA (2008) The social hyperlink. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 227-249.
- Adamic LA and Adar E (2003) Friends and neighbors on the Web. *Social Networks* 25(3): 211-230.
- Adamic LA and Huberman BA (2001) The Web's hidden order. *Commun. ACM* 44(9): 55-60.

- Albert R, Jeong H and Barabasi A (1999) Internet: Diameter of the World-Wide Web. *Nature* 401(6749): 130-131.
- Barabasi A (2003) *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, Cambridge, MA: Plume Books.
- Barnett GA, Chung CJ and Park HW (2010) Uncovering Transnational Hyperlink Patterns and Web-Mediated Contents: A New Approach Based on Cracking .com Domain. *Social Science Computer Review* 00: 1-16 (accessed 9 September 2010).
- Barnett GA and Sung E. (2005) Culture and the Structure of the International Hyperlink Network. *Journal of Computer-Mediated Communication* 11(1): 217-238.
- Bharat K, Chang B, Henzinger MR et al. (2001) Who Links to Whom: Mining Linkage between Web Sites. *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)*: 51-58.
- Blood R (2000) Weblogs: A History And Perspective. *Rebecca's Pocket*. Available at: http://www.rebeccablood.net/essays/weblog_history.html.
- Borodin A, Roberts GO, Rosenthal JS et al. (2001) Finding Authorities and Hubs From Link Structures on the World Wide Web. *10th annual international World Wide Web conference*. Hong Kong.
- Brin S and Page L (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1-7): 107-117.
- Broder A, Kumar R, Maghoul F et al. (2000) Graph structure in the Web. *Comput. Netw.* 33(1-6): 309-320.
- Brunn SD and Dodge M (2001) Mapping the "Worlds" of the World Wide Web: (Re)Structuring Global Commerce through Hyperlinks. *American Behavioral Scientist* 44(10): 1717-1739.
- Bruns A (2007) Methodologies for mapping the political blogosphere: An exploration using the IssueCrawler research tool. *First Monday* 12(5).

- Chang T, Himelboim I and Dong Dong. (2009) Open Global Networks, Closed International Flows: World System and Political Economy of Hyperlinks in Cyberspace. *International Communication Gazette* 71(3): 137-159.
- Dimitrova DV, Connolly-Ahern C, Williams AP et al. (2003) Hyperlinking as Gatekeeping: online newspaper coverage of the execution of an American terrorist. *Journalism Studies* 4(3): 401.
- Dimitrova DV and Neznanski M (2006) Online Journalism and the War in Cyberspace: A Comparison Between U.S. and International Newspapers. *Journal of Computer-Mediated Communication* 12(1): article 13.
- Etling B, Alexanyan K, Kelly J et al. (2010) Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization | Berkman Center. *Berkman center research publication* 2010-11: 1-46.
- Etling B, Kelly J, Faris R et al. (2009) Mapping the Arabic blogosphere: politics, culture and dissent. *Berkman center research publication* , 2009-06.
- Finkelstein S (2008) Google, links and popularity versus authority. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 104-120.
- Foust JC (2009) *Online Journalism: Principles and Practices of News for the Web*, Scottsdale, Ariz.: Holcomb Hathaway.
- Ghitalla F (2008) *L'atelier de cartographie. Pratique et enjeux des cartographies thématiques de documents web*. Available at: <http://www.webatlas.fr/download/docs/ateliercartographie.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=3e3ddf360a9c31c20dcbe3a3f678d2c6>.
- Ghitalla F (2009) *Du nuage aux abymes. Dimensions heuristique et expérimentale des modèles du web*. Available at: <http://www.webatlas.fr/download/DuNuageAuxAbymes.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=3e3ddf360a9c31c20dcbe3a3f678d2c6>.
- Ghitalla F, Le Berre A and Renault M (2005) *Des documents, des liens et des acteurs. Expérimentations autour de radiographies documentaires du web*. Conférence H2PTM.

- Halavais A (2000) National Borders on the World Wide Web. *New Media Society* 2(1): 7-28.
- Halavais A (2008) The hyperlink as organizing principle. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 39-55.
- Hargittai E (2008) The role of expertise in navigating links of influence. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 85-103.
- Hargittai E, Gallo J and Kane M (2008) Cross-Ideological Discussions among Conservative and Liberal Bloggers. *Public Choice* 134: 67-86.
- Harries G, Wilkinson D, Price L et al. (2004) Hyperlinks as a data source for science mapping. *Journal of Information Science* 30(5): 436 -447.
- Hespos T (2008) How hyperlink ought to change the advertising business. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 137-144.
- Hsu C and Park HW (2010) Sociology of Hyperlink Networks of Web 1.0, Web 2.0, and Twitter: A Case Study of South Korea. *Social Science Computer Review* 00: 1-15 (accessed 9 November 2010).
- Jacomy M and Ghitalla F (2007) *Méthodologies d'analyse de corpus en sciences humaines à l'aide du Navicrawler (Rapport final)*. Available at: http://www.webatlas.fr/download/methodo_shs_navicrawler.pdf?aa3b70196c0dc6fb4c4810f9d1c623ff=90d626e304d42bbf1fdb8397c44a569.
- Kelly J and Etling B (2008) Mapping Iran's online public: politics and culture in the persian blogosphere. *Berkman center research publication*.
- Kenney K, Gorelik A and Mwangi S (2000) Interactive features of online newspapers. *First Monday* 5 (1-3).
- Lin J, Halavais A and Zhang B (2007) The Blog Network in America: Blogs as Indicators of Relationships among US Cities. *Connections* 27(2): 15-23.
- Marlow C (2004) Audience, structure and authority in the weblog community. *Communication Research in the Public Interest (International Communication Association Conference)*, New Orleans.

- Marres N (2004) Tracing the trajectories of issues, and their democratic deficits, on the Web: The case of the Development Gateway and its doubles. *Information Technology & People* 17(2): 124-149.
- Napoli PN (2008) Hyperlinking and the forces of "massification". In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 56-69.
- Park HW (2003a) Hyperlink network analysis: a new method for the study of social structure on the web. *Connections* 25: 49-61.
- Park HW (2003b) Hyperlink network analysis: A new method for the study of social structure on the web. *Connections* 25(1): 49-61.
- Park HW (2010) Mapping the e-science landscape in South Korea using the webometrics method. *Journal of Computer-Mediated Communication* 15(2): 211-229.
- Park HW and Jankowski NW (2008) A hyperlink network analysis of citizen blogs in South Korean politics. *Javnost - The Public* , 15(2).
- Park HW, Kim C and Barnett GA (2004) Socio-Communicational Structure among Political Actors on the Web in South Korea. *New Media & Society* 6(3): 403 - 423.
- Park HW and Thelwall M (2003) Hyperlink analyses of the World Wide Web: a review. *Journal of Computer-Mediated Communication* 8(4).
- Park HW and Thelwall M (2008) Developing network indicators for ideological landscapes from the political blogosphere in South Korea. *Journal of Computer-Mediated Communication* 13(4): 856-879.
- Park HW, Thelwall M and Kluver R (2005) Political Hyperlinking in South Korea: Technical Indicators of Ideology and Content. *Sociological Research Online* 12(3).
- Quandt T (2008) (No) news on the world wide web. *Journalism Studies* 9(5): 717-738.
- Rogers R (2010) Mapping public web space with the Issuecrawler. In: Reber B and Brossard C (eds) *Digital cognitive technologies : epistemology and the knowledge economy*. London, Hoboken NJ: Wiley, 115-126.

- Rogers R and Ben-David A (2008) The Palestinian—Israeli peace process and transnational issue networks: the complicated place of the Israeli NGO. *New Media & Society* 10(3): 497 -528.
- Rogers R, Jansen F, Stevenson M et al. (2009) *Social Research with the Web*. Association for Progressive Communications and Hivos.
- Rogers R and Marres N (2000) Landscaping climate change: A mapping technique for understanding science and technology debates on the world wide web. *Public Understanding of Science* 9: 141-163.
- Schulman SL (2008) Hyperlinks and marketing insight. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 145-158.
- Scott J (2000) *Social Network Analysis: A Handbook*, London: SAGE Publications.
- Stray J (2010) Linking by the numbers: How news organizations are using links (or not). *Nieman Journalism Lab*. Available at: <http://www.niemanlab.org/2010/06/linking-by-the-numbers-how-news-organizations-are-using-links-or-not/>.
- Sunstein CR (2001) *Republic.Com*, Princeton: Princeton University Press.
- Tankard JW and Ban H (1998) Online newspapers: living up to the potential? AEJMC Conference, Baltimore.
- Thelwall M (2004) *Link analysis : an information science approach*, Amsterdam; Boston: Elsevier Academic Press.
- Thelwall M (2001) Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology* 52 (13): 1157-1168.
- Thelwall M (2003) What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research* , 8(3).
- Thelwall M (2006) Interpreting social science link analysis research: A theoretical framework. *J. Am. Soc. Inf. Sci. Technol.* 57(1): 60-68.

- Tremayne M (2004) The web of context : applying network theory to the use of hyperlinks in journalism on the web. *Journalism and Mass Communication Quarterly* 81(2): 237.
- Tremayne M (2005) News Websites as Gated Cybercommunities. *Convergence* 11(3): 28-39.
- Tremayne M, Zheng N, Lee JK et al. (2006) Issue Publics on the Web: Applying Network Theory to the War Blogosphere. *Journal of Computer-Mediated Communication* 12(1): 290-310.
- Tsui L (2008) The hyperlink in newspapers and blogs. *The hyperlinked society*. Ann Arbor: University of Michigan Press, 70-83.
- Turow J (2008) Introduction: On not taking the hyperlink for granted. In: Turow J and Tsui L (eds) *The hyperlinked society*. Ann Arbor: University of Michigan Press, 1-17.
- Turow J and Tsui L (2008) *The Hyperlinked Society: Questioning Connections in the Digital Age*, Ann Arbor.
- Watts DJ (2004a) The “New” Science of Networks. *Annual Review of Sociology* 30(1): 243-270.
- Watts DJ (2004b) *Six Degrees: The New Science of Networks*, New-York: Vintage.
- Wikipedia. Rickrolling - Wikipedia, the free encyclopedia. Available at: <http://en.wikipedia.org/wiki/Rickroll>.
- Williams AP, Trammell KD, Postelnicu M et al. (2005) Blogging and Hyperlinking: use of the Web to enhance viability during the 2004 US - campaign. *Journalism Studies* 6(2): 177.
- Zimmer M (2009) Renvois of the past, present and future: hyperlinks and the structuring of knowledge from the Encyclopedie to Web 2.0. *New Media Society* 11(1-2): 95-113.

¹ <http://socscibot.wlv.ac.uk/>

² <http://lexiurl.wlv.ac.uk/>

³ <http://voson.anu.edu.au/>

⁴ <http://www.issuecrawler.net/>

⁵ <https://addons.mozilla.org/en-US/firefox/addon/200263/>