Internet Research: The Question of Method

Richard Rogers

I'd like to introduce to you a research program that strives in some sense to redo the agenda for Internet research. I'm coming at this from a number of years of experience in building tools that try to embrace the methods in the media instead of importing methods from the social sciences and elsewhere. So that's my point of departure. I'm going to talk about where Internet research has been going, especially the major contributions made by the social sciences, in particular beginning around 1998, and then I'll move to the current period, introducing general approaches that I term "digital methods." Up until the call by Steve Jones in 1998, in his edited volume Doing Internet Research we were in an Internet as cyberspace period, where cybercultural studies dominated, with the idea of the Internet as a virtual realm apart, something that has an asterisk attached to it, something that is out there in its own world, with its own dynamics. Those cybercultural ideas have informed and continue to inform both popular as well as intellectual perspectives on what to do with the Internet. Is it a realm that allows for identity-altering transformation? Is it a realm that allows for different kinds of developments than the off-line? Should the Internet be studied separately? If it is studied as embedded in society, are user studies the only way to go about Internet research? How else to study the Internet for social and cultural research purposes?

It was in the year 2000 when the British ethnographers Slater and Miller came out with what to me was an important study. They grounded the Internet. They went to Trinidad and Tobago, and they studied Trinis' use of the Internet in cybercafés. And what they found was not that the Internet or that cyberspace was some kind of separate world apart, with those "inhabiting" it being transformed by it. Rather what the ethnographers found, which is of course typical of ethnography in general, was that the Internet was a space where Trinis performed their own culture. They appropriated the medium in ways which were Trini-specific. Whilst a case study, the implication of this work was more general. If the Trinis were doing it their way, mostly likely national or other cultures were doing it as well. In some sense it grounded the Internet both culturally as well as intellectually. But what I want to talk about today is what that sort of work accomplished methodologically.

Arguably it set a methodological agenda - you had to go off-line, you had to go to the off-line, or the ground, in order to study the online. One had to study users. And indeed this has been the social scientific project. To me some of the most significant work has been done by the research program run by the sociologist and science and technology studies scholar, Steve Woolgar, in what was called the Virtual Society? Program, from 1997 until 2002. The question mark was very important for them. They debunked first of all this idea of cyberspace as a realm apart. But they also subsequently grounded findings in a series of empirical studies. Woolgar formulated what he called "Five Rules of Virtuality." Among them are that there is no desolation for people who spent a lot of time online. Rather online activity stimulates more off-line activity. They

formulated what has come to be called the classic digital divide critique, which is to say that people's skills as well as the way people understand and experience the risks of the Internet are unequal. It has to do with particular demographics, et cetera. In formulating these rules, the program also solidified the dominant methodological program for Internet-related research in social science. The program has been summarized in the notion of "virtual methods." A couple volumes now have appeared, where the researchers continue to develop quite a classic social scientific armature, which includes interviews, surveys, observation, et cetera. What I would like to point out in particular is these could be categorized or conceptualized as digitized methods, that is, taking methods, existing methods, and trying to move them online. How best to do an online survey? Is Survey Monkey the way to go or not? Should one opt for the pro version? How best to formulate your first contact email with a group, with a community? All these sorts of things take into account some small differences that the online environment brings with it, and they make for slight changes to existing methods; so digitized methods with small amendments.

What I'd like to try to do - and I think that many of us in a sense are doing this already - is introduce a new era in Internet-related research where we no longer need to go off-line, or to digitize method, in order to study the online, or culture and society via the online. Rather in studying the online, we make and ground findings about society and culture with the Internet. So the Internet is a research site, where one can ground findings about reality. And with this particular idea I have introduced the term digital groundedness, or online groundedness, where claims about society are grounded in the online. I want to come directly to an example. Of course one of the seminal cases that has come out recently that I think could be situated under this term is Google Flu Trends. Google Flu Trends is different kind of Google project, because it's run by Google.org, the non-profit arm. Google Flu Trends uses the Web as an anticipatory medium, so it reintroduces the discourse that things happen online first, or you can find out what's happening in society first by going online. The online is quicker to the ground than other ways of getting to the ground. Google Flu Trends collates search engine queries, and geo-locates where these queries have taken place for flu - the word "flu," and flu-related symptoms - and arguably makes findings that are about one week ahead of those by the Center for Disease Control, which base their findings on emergency room reports and similar formal accounts. So the Web becomes an anticipatory medium again, and of course it's controversial to use the Web as the site to base claims about where flu is happening as opposed to on the ground, the emergency rooms, et cetera.

I want to give you one more example, which struck me initially. It comes from an article in a leading Dutch newspaper, the NRC Handelsblad, that came out in August 2007. They published a story where the question was: Is right-wing culture becoming more extremist in the Netherlands? This sort of question may be applicable for any number of countries, but what I want to talk about is their method. Instead of traditional investigative journalism, embedding a researcher (going native) or instead of going to a repository of leaflets and other ephemera - instead of using those standard methods, they used the Internet archive. They looked up in the

Internet archive about 100 sites. And they made a data set, an Excel sheet which they also published, in another special Internet-related data-sharing practice. They read the content of right-wing and right-wing extremist sites over a period of about ten years, and they found that the language on the websites over the years has become more and more extreme; the words were harsher and harsher. They thereby concluded that right-wing culture in the Netherlands is hardening. They made these findings on the basis of the Internet. Now for many people—who have in the backs of their minds that the Internet is a space apart, a virtual realm, or for those people who have a sense that the Internet may tell you something, but ultimately you have to go to the ground for your baseline, the newspaper's method and means of grounding claims are quite radical. They took the Web as the site to ground the findings about society.

So what I'd like to do today is think about what kind of data are available in the medium, first of all. And second of all I'd like to also think through the ways in which the Internet offers particular research possibilities. And by research possibilities, I'd like to introduce the question of learning from the methods in media. I would like to talk about what Internet-specific analysis would entail. I've already given you an introduction to it, but now I would like to take you through a series of digital objects, devices such as engines, platforms, et cetera and think through how they offer method. I would like to think through with you what I call "digital methods," repurposing methods in media for social and cultural research. I'd like to talk about the link, how links are normally studied and how I would propose that they could be studied. What kinds of opportunities are on offer if we follow the medium and its methods. I'd like to talk about the website, engines, and spheres. The blogosphere is of course the well-known one; scholars have coined the term websphere. I'd also like to talk about the newspshere. I view the spheres as engine-demarcated spaces. I'd like to talk about the Webs, in the plural. Who's the senator from Alaska with the Internets? Or was that George W. Bush? I think they were onto something. There are Webs in the sense of national Webs and much of this has to do with particular web technology that has emerged, GeoIP. I'd like to talk about in conclusion social networking sites and Wikipedia, how they are normally studied and how else one might study them if one seeks to learn from the methods of the medium and think about how to ground claims about society in those spaces.

The link. How are links normally studied? There are a couple of traditions. One is from humanities and hypertext theory. Links author potential stories; the surfer then becomes the author going from link to link, authoring. Similarly, in path theory, small-world theory, social network analysis, what we're concerned about often times is ideas of distance. So how far away are websites from one another, if we follow links -- building on Stanley Milgram's seminal work. And in particular that work is applied in trying to find the optimal paths. What are the optimal paths to reach someone or something? Social network analysis oftentimes concerns people and their positioning in networks. Are they central or are they in between? Highly in between? Et cetera. What would happen if we were to think through what to do with a link by following methods in the media? One would think immediately of Google. How does Google

treat links? Google treats links in some sense as reputation markers. That is to say, sites are ranked on the basis of the number of links they get and the number of links they get from highly influential sites. How can you make use of this particular way of thinking about links? I've created a piece of software called the Issue Crawler, which builds on the insight that links are reputation markers. What the Issue Crawler does is crawls sites, captures all the outlinks in any number of degrees of separation, and puts them into a dataset for different sorts of analyses. I'll show you two types of analyses. One is about the reputation of sites and what one can learn not from a sort of Google macro analysis treatment of the whole Web but rather a subset of the Web. That is to say, sets of linked Websites. One can profile an actor according to the links it gets and receives, either in total, or in a particular subject or issue area.

This is an example—classic work, I think this is from 1999. This was one of the first times I did this. It's an analysis of the micro-politics of association. It's about three different kinds of sites. You have the governmental sites in blue; commercial site in yellow; and an NGO in green. You'll notice Novartis links to Greenpeace, but Greenpeace does not link back. Both Greenpeace and Novartis link to government and government does not link back. These are classic politics of association. Governments normally only link to other governmental sites, et cetera. Normally NGO's don't necessarily want to endorse other sites that they're critical of by linking to them. This is increasingly the way links are made. In other words one can begin to understand or get a grasp of very normal politics of association that you can read from the online.

Another example. Here are three separate corporations, all in yellow. They're profiled according to the types of links they receive and the types of links they give. Their reputational status is different depending on the types of links they receive. If a commercial source receives links from government, it's a very different status marker than if it received links from only other commercial actors.

This is a picture of the Issue Crawler, just an example of the Issue Crawler output, in a particular subject area (e-culture in the Netherlands), with selected profiles of the top actors, and the links they receive and give, to the right.

I would like to talk a little bit about what else one can do with link analysis, in the area of Internet censorship research. I occasionally work together with the people at the University of Toronto in the Citizen Lab. It's the Open Net Initiative, which was on the cover of the New York Times two weekends ago for having discovered a cyber-espionage network, apparently allegedly operating out of China. They were contacted by the offices of the Dalai Lama because the Dalai Lama office's computers were acting up. And what they discovered and made public a couple weeks ago was an intriguing information warfare practice, social malware. I'm not going to talk about infowar; rather I'd like say something about my group's contribution to Internet censorship research, particularly methodologically. What the Open Net Initiative does is it makes a kind of directory of websites, with a number categories. Human rights sites, famous bloggers, humor sites, anonymizers—37 different categories. And in total, across all categories, at least when I

last undertook analysis, they have approximately 2,000 URLs; that's their sample. They use these 2,000 URLs and they query them, they fetch them, in each country in question to see the level of blockage, the level of Internet censorship across some 40 countries. However, I read in the Cyber Dissident Handbook that came out from the organization Reporters Without Borders, a Paris-based NGO, an article entitled the Worst Enemies of the Internet, or similar. There was a passage where the Saudi Arabia Information Ministry spokesperson boasted that they were blocking something on the order of 400,000 sites. And I said to myself, Open Net Initiative is checking 2,000 sites, the Saudis say they're blocking 400,000; how do we expand our lists? So I developed a technique called Dynamic URL Sampling whereby what we do is we take the initial list the Toronto researchers have drawn up, crawl all of the URLs and fetch all of the outlinks from these URLs, all these additional pages, check them against the original list and the ones that are left over we then subsequently check for blockage using initially some proxies but then later we run these through computers that are located on the ground in these various countries because of the unreliability or the intransparency of proxies.

What you see before you is a picture of a network of Iranian social, political, and religious sites—that's an ONI, Open Net Initiative, category for Iran. The ones in red are the ones that are blocked; the ones in blue are the ones that are not blocked; and the ones in red with the little yellow pins on them are sites that we discovered to be blocked. Those were previously unknown to be blocked, by the researchers. I'd like to highlight one site in particular, and the difference it makes when one uses hyperlink analysis over building a list of sites, in a manual practice. The ONI researchers had BBC.co.uk on their list, and the URL was resolving in Iran. When we ran the network analysis we also ran the page level and what we found was that BBC.co.uk/Persian was far more relevant than BBC.co.uk. And indeed the BBC.co.uk newspage wasn't blocked in Iran, but the Persian language BBC newspage was blocked. So we made a contribution not only to methodology, how to expand the lists through Dynamic URL Sampling, but also we made contributions to the findings. This stuff is double-edged. We would be very good censors. All those blue sites are waiting to be blocked. It's actually quite difficult to deal with this issue and the associates with the University of Toronto at the Berkman Center at Harvard sometimes talk about "data escrows" where they keep lists away from the prying eyes of censors.

The website—how is it normally studied? Well classically it's studied in usability circles. There's a debate between the "Don't make me think" school and those who are more interested in a poetics of navigation. Websites are often studied in design circles. You may or may not know that the majority of the Web is blue. Eye tracking is another classic method for Website study. The outputs are heat maps of eye movement, and these maps are useful. Websites are studied as something that needs to be optimized for any number of different reasons, largely because as I'll come to later in what I call the drama of search engine space. That is to say, you need to have your site in the top five, in the top ten, for as search engine user studies have found the number of pages and search engine results pages people look is in decline. Also and this is something that I've been hearing tiny bit here, people oftentimes study websites in terms of their

features, site features. So which sites have more features, and is there a correlation or relationship between the features of a website and the number of visitors? If there's more interactivity, is there more participation? Things like this.

This is an example of a heat map. This is the Google results page, where eyes are pointed upper left. They call that the golden triangle of search. And indeed, if you noticed, not too long ago Google moved its menu upper left.

How else to study it? I have been looking into the website as an archive object for some time now. If you run a quick search on Google Scholar and I've done this in the Web of Science as well, you'll notice that most of articles about the Internet archive are about how it works as opposed to how to use it. So what I have been trying to do is develop methods or means to use the Internet archive for research. Like learning from Google for link research, I follow the medium for clues or guidance. When you look at the Wayback Machine's interface at the Internet archive what you notice is that it privileges single site histories. You have to type in an individual URL, and you have returned to you a list of dates when that URL has been arrived, with an asterisk indicating that the URL content has changed. So what can you do with single site history? I mean first of all, would you like to study the history of the Web like that? Probably not. But there are things you can do with it. And so we've developed this tool or actually a couple of tools that can capture a website's history in order to tell a story about it. I made a video not too long ago, YouTube style video, and it concerns the Google Directory. You will be familiar with this particular project perhaps. The Google Directory sits on top of DMOZ, the Open Directory Project. It was put online by Google in 2000. And what has happened over the years is it's been marginalized. Now Google, the Google Directory and the human-edited Web more generally are in decline. I mean if you look at the seminal directory, the Yahoo! directory, now it's something that is based on an advertising model; something that you pay for to be listed in quickly. It is also no longer the default search engine on Yahoo!. On Google, the directory is no longer on the front page. What I wanted to do in this particular video is explore what one could learn from the history of a particular page as an example of how one might work with the Internet archive. The video is called Google and the Politics of Tabs.

> [This is the history of Google as seen through its Interface. From the beginning, sometime in November 1998 all the way up until late 2007. These are screen shots of Google Interface taken from the Wayback Machine of the Internet Archive. The history of the Google is important. For some people, Google is the Internet. And for many, it's the first point of access. And Google, as the face of the Internet, has remained virtually the same over the past ten years. But there have been some subtle changes to the Interface. So let's go back and look at this in a little bit more detail. You see initially Google with a standard Web search button and its intriguing "I'm feeling lucky button" have been your only options. Then the Directory gets introduced with some front page fanfare. It's the Open Directory Project, DMOZ.org, that Google's built an engine on top of. Then come the Tabs on top of the Search box with the Web search being privileged at

the far left, followed by Images, Groups (that's searching Usenet), and the Directory makes it to the front page. News, the Google news service, the news aggregator was next. Froogle is introduced; that was that cost comparison e-commerce service. And that stayed on the front page for a while, then was dropped. Followed by Local, which later became Google Maps. You can see that the services are becoming more and more present; there are now five or six at the top bar. Then they add a "More" button. What we're interested in is which services remain on the front page and which get relegated to "More" or "Even More." But let's look at this in some more detail. Let's look at the fate of the Directory over time. It's a story of the demise of the librarian, of the demise of the human editors of the Web, and the rise of the back end, of the algorithm taking over from the editors. Now you see that it's introduced with great fanfare in 2000. The Web is organized by human editors. It remains on the front page. It achieves the Tabs status that we talked about previously. Fourth Tab here. And keeps its place on the front page even as other services are introduced. However, in 2004 something happened: It got placed under the "More" button. You had to click "More" to find the Directory. And in 2006, if you clicked "More," the Directory wasn't there; you had to click "Even More" and there you would find the Directory. As it loses its standing, it also loses recognition. Lots of people don't really remember that there is a Directory just like other services that have left the front page real estate. Also of interest are the services that climb from being "Even More" to "More" and all the way to the front page. But with the Directory, it's a sadder story. As the interface of Google moves upper left, and you click "More," you see that there's no Directory any longer. And you also see that there is no "Even More." So nowadays you have to search Google for its Directory to find the Google Directory.]

Okay, the "Even More" button is back by the way. It wasn't there at the end of 2007, and now it's back under the more button on the upper left menu. In the film I talk about the rise of the back-end, the rise of the algorithm, and the demise of the human. I would like to point out something important. If you go to the Yahoo! directory, and you type in a query, what you get back is a listing of sources which are now ranked by popularity by default. That is to say the sources are no longer in alphabetical order. So the egalitarian listing of information sources is no longer the default at the major directory. The algorithm is also spreading beyond the Internet to other digital spaces. In a comparative media analysis perspective you would look into these sorts of new rankings and recommendations over the alphabetical list; think of TiVo.

The engine—how are engines normally studied? There's a body of work on the politics of search engines that looks into search engines as sites or spaces of inclusion or exclusion, where particular sites are buried or they no longer exist in practice if they're way down in the rankings, or certainly out of the top 1000. Engines are also studied in terms of what could be called the attention deficit disorder. Users are increasingly looking at fewer and fewer returns, fewer and fewer pages. Jansen and Spink in particular have been studying this for a number of years and not only are people looking at fewer and fewer pages and returns but they're clicking sites that

are close to the top more and more often. So engines are studied as a space where placement really matters. They are also studied in terms of the notion of Googlization. It's a term that has been introduced or been worked on by library science scholars in particular as a reaction to the Google Books project. The minute Google entered the hallowed halls of the library, library science scholars began critiquing Google quite heavily, but also developing sophisticated ideas about what Googlization would imply for knowledge provision, and knowledge access more generally if it keeps going like this. Googlization arguably as a term has a political economy connotation. Google is creeping into more and more different services. They're no longer just a search engine; that much is obvious. Google and engines are also studied from surveillance and privacy studies points of view. In particular search engine results are being personalized on the basis of your histories. If you are signed into Google in particular, queries are not only logged but results personalized. And it's interesting in my view that it's becoming more and more difficult to study Google results, because Google results are not necessarily the results increasingly of some "one" algorithm but they're also partly your results. People increasingly do not receive the same results. So I call Google the "inculpable engine," as it's taking itself off the hook by having the user influence the results. But before it did so, or while it's doing so only a bit, what I have been developing is a means to study engines—Google in particular—as an ordering device, as an epistemological machine. And in order to do that, I have captured and stored the engine results, which is not in compliance with Google's Terms of Service. So I put up a notice asking for its forbearance. I look into what is an understudied aspect of engines, that is the volatility or stability of the actual results. Do results change day by day or are they relatively stable? Does it matter *when* you search for the kind of results that you receive?

Here is one example. These are the results of a query, made daily, over a 30-day period, November 2007. It's a query for RFID and within a 30-month period you see that most of the sites are rather stable. But some returns vary somewhat from a top ranking of four to the low ranking of 12, from the top ranking of 11 to 26, from 1 to 17, and from 14 to 31—that is some change, also given how users interact with results. The major change is one particular site that went from rank number 9 to 213 during the 30-day period.

What kinds of questions may be asked, when working captured search engine results? What I was interested in was a follow-up on the classic idea of what social research is according to C. Wright Mills. It is to present conflicting realities. And what I was looking at was whether or not Google results are increasingly becoming more and more familiar, that is, able or unable to present conflicting realities. The question is whether the results that come out of Google are aligned with the familiar or the mainstream. That is, are Google results becoming quite similar to the sources that you would hear on the evening news? So I've been tracking results of the query for 9/11 over about two years, collecting the top 1000 results for that query because Google serves a maximum of 1000 results, and looking in particular at the rankings over time of three important sources for 9/11 accounts generally: the New York Times, the New York City government, and a third one which I'll tell you about in a second. The New York Times is in

blue; the New York City government, nyc.gov, is in green. The third one is a site that presents a conflicting view of reality, the Movement for 9/11 Truth, 911truth.org. For approximately six months 911truth.org was in the top five results of Google. Something happened around the 17th of September 2007; it dropped precipitously from result 5 to result 200 and then off the charts to under 1000. I believe this is the first fully documented case of the disappearance if you will or the apparent removal of a site from Google results. There are a number of reasons why this may have occurred which I'll get into later in the Q&A.

In the case above, we are studying Google mainly, though one could interpret 911truth.org's high placement as socially significant. But how to use Google to study more specifically what's happening in society? We've built a piece of software called the Google Scraper, also known as the Lippmannian device because Walter Lippmann was always interested in equipment or tools that could provide a "coarse view" of the partisanship of an actor. So let me just show you what one is able to find with the Lippmannian device. What we do is we capture engine results from one query. In this particular case, the first 100 results for the query climate change. Then for each of these 100 sources we query them individually for a particular subject matter. In this particular case, for the names of climate change skeptics. We're interested in seeing whether or not we can provide a coarse sense of the partisanship of a particular organization. These are the top 100 returns for the query "climate change" in Google in July 2007. We've done this for other periods as well, this just being a snapshot. We queried each of the individual sites - the EPA, BBC News, UN Environmental Program, IPCC, Pew Climate, et cetera, et cetera, for the names of well-known climate change skeptics. We wanted to look into not only whether we could detect the partisanship of the source but also we wanted to look whether the Web is like the news, providing quite a lot of space or voice to the skeptics, or at least to climate change skepticism. What you see here are the results. You see that the skeptics actually are not named too often on very many sites. What jumped out for me on this particular one is climatescience.gov actually names skeptics. Marshall.org often times funds them; sourcewatch.org is a watchdog organization. With the Lippmannian Device you gain a sense of not only partisanship, but also issue or position commitment per source.

The spheres. How are blogs often studied? Blogs are often studied as a genre; they're recognizable because they have particular formats: reverse chronological order, a blogroll, et cetera. Blogospheres often times are studied in relationship to the news. What other researchers reported earlier on today is counterintuitive to me; in the previous studies that I've looked at and the studies that I've done, I've always found the blogosphere to be parasitic on the news as opposed to the news being parasitic on it. But in any case the blogosphere is often quite obsessed with mainstream media. Blogs are also studied as organizers of voice, voice-giving, or as authentic voices as in the case of the famous Iraqi bloggers, et cetera.

How else to study spheres? I take spheres to be engine-demarcated spaces. That is to say the blogosphere, as sphere, is in some sense authored or at least demarcated by Technorati. I take the Web as in some sense demarcated by Google; I take the newssphere if you will as demarcated by

Google News; I take the social group marking sphere, or tagosphere, as demarcated by Delicious. And what I then do is perform "cross-spherical" analysis. That is to say, what are the differences in available or privileged sources between these subspaces on the Web? I also ask questions about the quality of media. Is the blogosphere something that treats issues, subject matters, substance differently, qualitatively differently, than the news or the Web? One brief case study, again having to do with climate change, is called the Issue Animals project. We queried the various sphere-engines for climate change, and saved the results. We subsequently made a list of animals associated with climate change, manually, from reliable sources. We queried each of the sources per sphere for these animals, that is on the Web through Google Web search, in the news through Google News, and in the blogosphere through Technorati. We queried all of these individual sources to see whether or not particular animals are privileged per sphere, in particular to look into whether each of the spheres have tendencies towards creating media icons or not. When looking at the news, we noticed that the polar bear really stands out; when we look at the blogosphere as I said what I have found it to be quite parasitic on the news, the news becoming amplified in the blogosphere. On the Web intriguingly the animals are treated in a more equal fashion. That may say something about the "quality" of the Web over the news and over the blogosphere, or at least its lacking of media icons.

The Webs. This is in the plural as I mentioned. How are they often studied? They're often studied in the singular, the World Wide Web, or Web Studies. They're often times studied as I mentioned before in terms of cyberspace. Also as a technical infrastructure, as a technical infrastructure with particular end-to-end principles and then with particular engineering, the packet-switching, which conventionally at least way back when was supposed to allow us to root around censorship. Anyway, cyberspace, the idea of cyberspace arguably grows out of a particular understanding of the effects of the infrastructure. And the Web is also studied as a realm apart, as something different, certainly not as a potential baseline.

What I started to do two years ago was to see if I could use the Web as an indicator for the state of a country. I was exploring Iraq in particular. To know what was happening in Iraq, we had news reports, we had the authentic bloggers from Iraq, we had some Presidential candidates and Congress people that were on fact-finding missions; there were no tourists until very recently, actually you may have read that in the New York Times I think it was or International Herald Tribune, the first Western tourists have arrived in Iraq; there's always pilgrims moving through there, too.

So what I've tried to do is develop a means by which I could find out what was going on in Iraq by looking at the state of their Web. How broken is it? Were the university websites up? Things like this. So on the basis of this, I began to develop a series of methods in order to diagnose the condition of a society according to its Web.

The reason for national Web thinking more generally, more conceptually, has to do with GeoIP technology and the idea that the Web has been grounded—you may have noticed it; I notice it a

lot because I'm a baseball fan and I subscribe to MLB.tv and any time I'm in the US I'm often blacked out, whereas when I'm in Europe I can watch all the games. You'll have noticed this with the Olympics. The content is served according to your location, and that is, your IP address is detected. This technology goes back to the famous Yahoo! Lawsuit brought by two French NGO's in 2000, where the case was about blocking French users from looking at Nazi memorabilia pages in the US. The technology, GeoIP location technology, was developed directly as a result of that particular lawsuit. Goldsmith and Wu have written a book about it.

So what kinds of ideas can one gain about a society if you look into the condition of its Web? Here are some ways to study the condition of the Web; I'll just list certain Web-diagnostics briefly. Youthfulness. Are pages fresh? You do that through analyzing page date stamps. How broken is the Web? You can use link validators. To find out the cohesiveness of a national Web, you can use hyperlink analysis; are sites linking to one another nationally, or more internationally? How gated is the national Web? Also, how dated are the users, if you will? That is to say, what kind of browser versions are they using? You can use server logs for that.

Here is another way to gain indications of societal compositions according to the Web. This is a top and second-level domain topology of the US, which shows a ranking of which domains are in use. You can see that ".com" is the most important by far in the US. Compare that to Palestinian territories, where ".org" is most significant.

I'll briefly mention the Palestinian Web mapping project, in this context, for it demonstrates how to study social divisions. We did this study with Cambridge University and the University of Toronto. It is an effort to see what's going on in the Palestinian territories by looking at the Web. We took all the Fatah-related websites and all the Hamas-related websites, and we looked at their linking patterns. What we found is that Fatah sites link to the news, to local NGO's, to international NGO's, and Hamas was linked to no one. They link only to RSS readers, which gives you an indication of how their networks operate. The work also provides a particular view of the differences between Fatah's and Hamas's information cultures, one linking with national, international NGO's, news organizations, et cetera, and the other off-the-radar, with individuals receiving the information through subscription only. And it also gives you an indication of the splits on the ground, two different kinds of Web presences and cultures.

Social networking sites. How are they normally studied? They're often studied in terms of Goffman's idea of the presentation of self. One of the more intriguing reports was about social class divides performed on social networking sites. MySpace and Facebook are said to have different classes of users. The extent to which Facebook is a walled garden versus MySpace is another way of looking at it. Social Networking Sites are also studied in terms of the difference or similarity between real-life friends and friended friends. More significantly perhaps is how difficult it is to defriend online, how that amplifies the effect. Any kind of touchy social relations are not resolved very well by clicking, and having alerts broadcasted.

How else can they be studied? I've been introducing a term recently that I like to call "post-demographics." The term takes into account the kind of information on the profile of social networking sites that is different from the standard demographics. What's on a profile of a social networking site? The ones that are highlighted are favorite media interests. That is to say, movies, TV shows, books, heroes, et cetera. So with post-demographics I propose to again follow the media and study how profilers already make use of these preferences, of these particular favorites. Then I'd like to repurpose the way they do it, their method, for social and cultural research.

This is my first attempt at it; it's more of an art project. It's called elfriendo.com. Check it out. I made ElFriendo when my team and I were Artists in Residence at the National Media Arts Institute in the Netherlands, Montevideo. It gives an indication of the work that can be done in post-demographics. It shows the the difference in favorites between all the friends of Obama versus all the friends of McCain. What are their friends' aggregate favorites? One could study Obama's and McCain's supporters, according to demographics. But what about studying Obama's and McCain's friends, according to their favorites? It's interesting that they have quite distinctive profiles, according to the music that the friends listen to, the movies, TV shows, books, heroes, et cetera. And TV shows may be of interest to advertisers and political consultants. Obama's friends' favorites: The Daily Show, Lost, The Office. McCain's friends: Family Guy, Project Runway, Top Chef, America's Top Model, Desperate Housewives. The larger point of post-demographics is that relationship between candidates and friends' media favorites may be distinguished from the relationship between candidates and supporters' demographics (gender, income, education level, et cetera).

Wikipedia. This is the last one. How is it often studied? It's often studied in terms of its accuracy. You will have seen the studies in Nature about Wikipedia vis-a-vis Encyclopedia Britannica. It's also studied in terms of its "encyclopedia-ness" if you will. Indeed, it is remarkable that Wikipedia is encyclopedia-like. It's also often studied or often used and studied as a kind of scandal machine. This practice in particular has picked up since Virgil Griffith at Caltech made the Wiki-scanner, which de-anonymizes anonymous edits. And it also studied or thought about at least in terms of the highly "vigilant" community. How can Wikipedians be so vigilant and also accurate given the fact that they're a) amateurs, b) free labor, et cetera? To put them to the test, what people have done - and there are a number of scholars have done this and later regretted it—at least that's what they wrote—they create false Wikipedia entries or they change things in a Wikipedia entry and then wait for something to happen. What happened was that many of these changes were corrected quickly, which came as a surprise. Wikipedians are highly vigilant. How?

I just want to tell you first that the Wikiscanner rocked the Netherlands. One of the Princes, Prince Friso and his Princess apparently were caught editing a Wikipedia entry. On the entry, Princess Mabel's, it was written that she had given "false and incomplete" information to the government prior to wedding the Prince. And this was in quotation marks in the Wikipedia entry.

It was found that the Royal Family, or at least their IP address, had removed one of the words, changing, “false and incomplete” to “incomplete.” It was front-page news, and created a scandal. What wasn’t reported was that the edit was reverted (changed back) within minutes, because one of the vigilant Wikipedians probably received a software alert saying that the entry had been edited.

One of the things that I’d like to point out is that most Wikipedia research today has forgotten the bots. In fact, if you go to the statistics of Wikipedia, the top Wikipedians are bots; they’re not humans. The bots are working in tandem with the humans. Why are Wikipedians vigilant? They have bots. And they have software alerts that tell them when something is changed, something has been reverted, or that something’s been edited, et cetera.

So the initial question is, how dependent is Wikipedia actually on bots? And how would one begin to think that through? It turns out that in total the number of human edits is far greater than the number of bot edits. However, when you look into different languages, individual language Wikipedias, you see that particular languages are more reliant on bots than others. In particular the languages that are most endangered have the most bot activity over human activity. What the bots are doing? They’re looking for vandalism, they are interlinking pages, et cetera. The most active bots are referred to internally as maintenance bots, a term which is disarming. However, the question is, where does the ‘maintenance’ end and content co-authorship between humans and bots actually begin?

Thank you very much.

[applause]

Dr. Shulman: All right, thank you Richard as always for telling us what everybody’s doing and has done in an authoritative way and for telling us all the things that we might consider doing, which is one of the reasons that we’re here. So it’s question time for Richard. I’ve got a hand in the back, Steve?

Steve: Okay Steven [name] at Cornell University. My team, the research group that I’m a member of at Cornell, studies very similar things to the things you talked about today. One of the things we do is take advantage of problems in the algorithms to make the algorithms better and let me give you an example. If you flip back to your slides to the section on where you were pulling down the hundred search results and then querying the next level of detail to look for under-sited authors. We’ve actually built systems that do essentially the same thing but in order to improve rankings based upon user preferences. So what you think about the fact if we do that, any implications of it?

I think it’s excellent. Let me contextualize a larger point, and then I’ll come back more specifically to your question. Recently one of the leading new media theorists, Lev Manovich, has called for a program called Cultural Analytics. The term is borrowed from Google Analytics.

It would build quite large-scale data collection facilities to take advantage of all the digital traces online and analyze them to think about culture production, state of culture, et cetera. That's a particularly large-scale model; it's kind of big science type of idea. And what I'm interested in are far more modest research undertakings. That is, instead of thinking about the models of these large companies and their large datasets and getting negotiated access to them, I'd rather think about ways that we can use the methods and computing techniques that are being implemented online and then think through what kinds of other sorts of research can be done with them, how these sorts of techniques can be repurposed. So indeed when you create techniques to better the rankings, or improve algorithms for ranking, I'm interested in using those algorithms—repurposing them—in order to query different sites to tell us whether or not these sites are in league with a particular position or friendly with particular kind of funders, et cetera. So it's a different kind of purpose but we build on the very important work that you're doing.

Questioner: There's a huge level of funding that's being pumped into this area, it's called, we call it "learning to rank" in the computer science world. And so thoughts that you have about this area would be very useful to any team that's working on that research agenda.

Thank you.

Dan: Hi, I'm Dan [name] at the Fletcher School. I'm interested in your discussion of sort of national Web sort of diagnostics. And the focus on Iraq is fascinating, I mean Iraq is certainly an extreme case. To what extent have you gone beyond Iraq, I mean to what extent have you developed metrics to try to and how much data have you gathered on all 178 or whatever you know it's embarrassing as an IR purpose to say I don't know the actual number of jurisdictions but you know the number of (CCTLD's?)?

245. No, I'm not sure how many. I mean, we're now developing metrics. So we've done a very brief case study on Iraq. We've done a more extensive study on Palestine. And of course these are very specific.

Dan: My question is they are outlier cases. So if you want to apply, I mean you're gonna get interesting stuff from that but if you want to try to develop a metric that's gonna apply more broadly.

Sure. Let me just address a larger point and then I'll come to the specifics. I make situated software. So what I normally do is develop techniques, software applications, for specific kinds of research questions. And then later I see whether or not they could be made into something more generic. So this is a particular kind of research practice and I just want to make this clear. It's very different from the standard social scientific instrument-making whereby you build an all-purpose tool that you then install on your machine, then you go and look for datasets, and plug them in. So what I do is I normally make things that are for quite specific situations. But in most of the work that I've shown you apart from the Iraq case, many tools later have been developed in to something more generic, that is for more than just the one research project. What

I'm trying to develop at least in the first instance are a set of metrics that will aid in diagnosing the condition of particular kinds of country when we don't have good diagnoses. Initially, at least this was the idea with Iraq and this is also the idea in the Palestinian case. I mean someone wrote me very recently and wanted to do something on Zimbabwe. But, apart from these specific cases, what can we learn from more country data? For example, from the Alexa data. I mean have you looked at the top 100 sites per country? It's very interesting. You can profile a country according to what kind of sites are in that top 100. Which kinds of countries are relying on the mega-upload sites? Just to give you one short example. So one can think about different sorts of Web indicators for ideas about the societal condition.

Dr. Shulman: Charli.

Charli: Yeah, I like your discussion of links. Actually I like the way you organized it from how it's been studied and what might be done differently. But on links, I was interested in the contrast you were drawing and I wasn't sure I understood the contrast between social network analysis which was in the "how it's been done before" versus this sort of reputational understanding, which is the "how might we do it." How is looking at sorts of links as reputational different from just in degree centrality as a prestige measure as social network analysis?

Yeah, that's a very good question. What I tried to do is go one step further and try to talk about the micro-politics of association. So that would be the answer in short. So normally when one would study hyperlink relationships in a qualitative way, one would try to think about whether or not they have an off-line relationship, whether or not they're partners, whether or not they're allied—you know all these sorts of things. And that's how one would explain why it is that they're linked. That's how one would do it in a kind of social networking sense, if you will. But what I'm proposing is you can find a politics of association where you don't need that kind of baseline of off-line relationships in order to come up with a reputational marker.

Ken: Ken [name]. I'm so grateful that you're doing things like this. One of the things that happens to me is I immediately start thinking: What kinds of interesting questions can we ask and then use these tools to answer? And I would like to say, what are two or three interesting research questions that you think these would answer. And a follow-up on Dan's question about Iraq and Palestine—I just think the next one is China and I'll just give you a quick answer. We have a group that comes regularly and does some executive ed from China and I give them some lectures on technology and technology-policy and we have this conversation about "censorship" of the Internet in China. And they acknowledge that these websites are blocked but they are grateful to their government because it prevents terrorism. And so I'd be really interested in having some of this evidence and say, "Okay so this is blocked but what kind of things does it block?" Just throw that out, I mean my brain's going a little faster than I think I can articulate, but what interesting broader research questions would you like to apply these things to?

One of the things that I'm interested in in the context of Internet censorship research is this: To what extent does a circulationist Web sort of pre-empt or forestall state Internet censorship? I didn't show it, but I did a case study on the Balochi or Baluchi Web, some people pronounce it different ways—sites in Pakistan. Those sites are routinely blocked by the state. However, the question is: Does one see the Web in a kind of old media style, as a set of single websites that are blocked or unblocked? Or do you see the Web as a content circulation space? So what I've tried to do is test the idea that the Web is a content circulation space by taking all of the sites that are blocked in Pakistan—all the Baluchi sites—skimming the content off of them, and then querying engines for that content to see whether or not, literally, that content has been repackaged, or moved to other parts of the Web. The answer is: very little. Pakistan is doing a very good job censoring Baluchi sites, and Baluchi content is not circulating well by other means. I did a similar project on China. I looked into the women's rights case. China routinely blocks at least three or four women's rights sites, according to the Open Net Initiative. I looked at what issues are discussed on those sites. The one child policy. The high suicide rate amongst women in China. The issue of sexual diseases. All kinds of sensitive subjects for the government. Some nine of them; I can't list them all straight off the top of my head at the moment. In any case, I wanted to see whether or not that kind of content was available on women's rights sites that aren't censored by China. Answer: yes. I could share this with you, so you could show them.