Access
to this site is Denied

برابر قوانین جمهوری اسلامی ایران و دستور مقامات قضایی
دسترسی به این سایت مجاز نمی باشد.

National Web Studies:
Mapping Iran Online

Richard Rogers, Esther Weltevrede, Sabine Niederer and Erik Borra
February 2012

Iran Media Program gov com org UNIVERSITY OF AMSTERDAM digital methods initiative

Acknowledgments

About the authors

Richard Rogers, PhD holds the Chair and is University Professor in New Media & Digital Culture at the University of Amsterdam. Esther Weltevrede, Erik Borra and Sabine Niederer are pursuing PhDs in New Media & Digital Culture, University of Amsterdam.

Cover image: Block page by Iranian ISP, Pars Online, circa 2009. The current block page is included in the appendix.

Abstract

This work offers an approach to conceptualizing, demarcating and analyzing a national web. Instead of defining a priori the types of websites to be included in a national web, the approach put forward here makes use of web devices (platforms and engines) that purport to provide (ranked) lists of URLs relevant to a particular country. Once gathered in such a manner, the websites are studied for their properties, following certain of the common measures (such as responsiveness and page age), and repurposing them to speak in terms of the health of a national web: Are sites lively, or neglected? The case study in question is Iran, which is special for the degree of Internet censorship undertaken by the state. Despite the widespread censorship, we have found a highly responsive Iranian web. We also report on the relationship between blockage, responsiveness and freshness, i.e., whether blocked sites are still up, and also whether they have been recently updated. Blocked yet blogging portions of the Iranian web show strong indications of an active Internet censorship circumvention culture. In seeking to answer, additionally, whether censorship has killed content, a textual analysis shows continued use of language considered critical by the regime, thereby indicating a dearth of self-censorship, at least for websites that are recommended by the leading Iranian platform, Balatarin. The study concludes with the implications of the approach put forward for national web studies, including a description of the benefits of a national web health index.

Introduction:
National web studies

In 2007, Ricardo Baeza-Yates and colleagues at Yahoo! Research in Barcelona published a review article on characterizations of national web domains, where they sketched an emerging field, which we would like to call national web studies. Of particular interest in the article is the distinction the authors made between studies in the 1990s on the characteristics of *the web* to those a decade later on *national webs* (Kehoe et al., 1999; Baeza-Yates et al., 2007). The term national web, we feel, is useful for capturing a historical shift in the study of the Internet, and especially how the web's location-awareness repositions the Internet as object of study. A national web is one means of summing up the transition of the Internet from "cyberspace," which invokes a placeless space of email and packets, to the web of identifiable national domains (.de, .fr, .gr, etc.) as well

as websites whose contents, advertisements and language are matched to one's location. The notion of the national web, we argue, is also worthwhile beyond the conceptual. It enables the study of the current conditions of a web space demarcated along national lines, as Baeza-Yates and colleagues pointed out in comparing one national web with another. As we would like to pursue here, it also may be useful for the study of conditions not only of the online, but also of the ground. That is to say, national web studies are another example of country profiling.

Here, building upon the web characterization work, we provide an approach to the study of national webs that provides an overall rationale for their study (why study a national web) and engages a series of methodological debates (how to study a national web).Where the latter is concerned, we put forward an approach that is cognizant of the multiplicity of user experiences of the web as well as the concomitant web data collection practices that users may actively or passively participate in. Search engines and other web information companies such as Alexa routinely collect data from users who search and use their toolbars, for example. Platforms where "crowds share" by posting and by rating are also data collection vessels and analysis machines. The outcomes of these data gathering and counting exercises are often ranked lists of URLs, recommended to users. When location is added as a variable, the URL lists may be country or region specific. The same holds for language; namely, websites are served that are in whole or in part in a particular language. Thus, in practice one is able to speak of country-specific and/or language-specific webs organized by the data collected and analyzed by engines, platforms and other online devices. There is a caveat: users of these devices draw upon their own data, and are recursively provided a selection of considered URLs. Such personalization may influence the country and language-specific URLs served however, to date the impact on search engine results appears to be minimal (Feuz et al., 2011). Consequently, the effects of personalization are not treated here (Pariser, 2011).

We term the interaction between user and engine, the data that are collected, how they are analyzed, and ultimately the URL recommendations that result, "device cultures." In the case study below, we discuss a series of device cultures and the kinds of national webs they organize. We discuss a blogger's, an advertiser's, a surfer's, a searcher's and a crowd's web, each formed

by the online devices and platforms that collect their data and ultimately purport to represent or provide in one manner or another a country-specific and/or language-specific web. Put differently, we are making use of web devices that "go local," i.e., devices that not only collect but serve web content territorially (which is usually nationally) or to a particular language group. In certain cases the two distinct meanings of going local are reconcilable; in other cases they are not. An engine may serve language-specific websites originating from inside the country as well as from outside the country in question. For example, in return for a query, the Bolivian "local domain Google" (Google.com.bo) may just as well serve results from Spain or Colombia as from Bolivia, with all being in Spanish. Thus when discussing the demise of cyberspace, and the rise of a location-aware web, there is a tension between two new dominant ways of interpreting the object of study: national webs versus language webs. We are sensitive to the tension between the two new manners of approaching the web after cyberspace, and are aware that "the local," which as mentioned above is how Google terms its national domain engines, may refer to either a national web, a language web or both.

We also discuss approaches to demarcating a national web, including sampling procedures. We are particularly interested in the fruitfulness of research outcomes from both keeping separate as well as triangulating the various parts of a national web—the blogger's, the advertiser's, etc. Are the URLs that are listed as "top blogs" by blog aggregators similar to the URLs that are listed as interesting by crowd-sourcing platforms? Does the list of URLs with high traffic, and available advertising space for speakers of a particular language (i.e., Persian), resemble that of the most visited websites in a related country in question (Iran)? We conclude that keeping the parts of the web and the lists of URLs separate may be beneficial, as a national blogosphere may have different characteristics than a national crowd-sourced web.[1]

Where the overall rationale for studying a national web is concerned, it implies not only a critique of the web as placeless space, and as universalized; it is also a means to develop further analyses of relationships between web metrics and ground indicators. That is, another aim of this study is the consideration of digital methods to understand the significance of national web space. By digital methods we mean algorithms and other counting techniques whose inputs are digital objects, such as links and website response codes, and whose application pertains to, but ultimately moves beyond, the study of online culture only (Rogers, 2009). We discuss metrics for analyzing the health of a national web, such as its responsiveness, freshness and accessibility. We have experimented with such analyses before, seeking to diagnose the condition of Iraq (in 2007, some four years into the Iraq War) by looking at "its web." We found a broken web. Iraqi university websites were down, or had their domains poached and parked. Iraqi governmental sites were suffering from neglect, with the exception of the Ministry of Oil (oil.gov.iq), which was bilingual and regularly updated. In our brief foray into the state of Iraq via the Iraqi web, we sought to develop a series of metrics for diagnosing the health of a web, which are both conceptual as well as empirical.

Blocked yet blogging: The special case of Iran

The case study in question is Iran.[2] It is in many respects a special case, not least because the term national web itself may be interpreted to mean the separate Internet-like infrastructure that is being built there (Rhoads and Fassihi, 2011). It is also a special case for the scale and scope of Internet censorship undertaken by the state, which is coupled with the repression and silencing of voices critical of the regime. In other words, the Iranian web is experienced differently inside Iran than it is outside of Iran, which is of course the case for all countries where state Internet censorship occurs. It is also seemingly authored differently from outside than from inside Iran. As a consequence many Iranians online, either site visitors or authors, whether inside or outside the country, need to cope with censorship. Inside the country, coping could mean being frustrated by it, and waiting for a friend or relative to bring news about a VPN or another means of getting around blockages. It could mean routinely circumventing censorship through VPNs, proxies, Google Reader and other means. Both inside and outside the country, coping may mean actively learning about (and consciously not using) banned words, and perhaps em-

---

1 The data for this study are online at the project website, http://mappingiranonline.digitalmethods.net/.

2 According to the International Telecommunication Union, 13% of the Iranian population uses the Internet and 21% of Iranian households have Internet access (2011). The marketing research reports an urban concentration of users, with "the vast majority (being) young, mostly 15 to 40" years of age (NetBina, 2010: 10). Figures on the Iranian diaspora are not available.

ploying code words and misspellings instead. It could mean self-censorship. The degree to which Iranians online express themselves in times of censorship is of interest here. Dealing with online thuggery is another matter, which we are aware of, but do not cover in any detail. For example, one may be warned or pursued by the Iranian cyber army (Deibert and Rohozinski, 2010). One copes, or protects oneself, through the careful selection of one piece of software or platform over another, based on which one provides safeguards and forms of anonymity. One may use wordpress.com for the ease with which one may choose a new email address as a login, or Friendfeed for the capacity to change usernames.

Having mentioned some of the reasons why it is a special case, we also would like to point out that certain general metrics such as site responsiveness and freshness may be put to good use when studying countries such as Iran. For example, if sites are blocked by the state, yet still responding and updated, one may have indications of a reading audience, both outside but also inside Iran. One may have indications of widespread censorship circumvention, as we report. Here in particular the retention of the separate webs in our sampling procedure is beneficial. That is, the Iranian blogosphere, or the Iranian bloggers read through Google Reader and indexed by Likekhor (a website that rates websites by 'likes'), are roundly blocked by the state, yet remain blogging. "Blocked yet blogging" may be the catchphrase for at least certain vital parts of the Iranian web.

Perhaps not often recognized as such, national webs are nevertheless routinely created. It may be said that national webs come into being through the advent of geo-location technology, whereby national (or language) versions of web applications (such as Google) are served nationally (Google.gr for Greece) together with the advertisements targeted to locals and information in compliance with national laws (Goldsmith and Wu, 2006; Schmidt, 2009). Notably, it is the search engine whose mission statement is universal access that is at the forefront of the rise of the national and the demise of placelessness (Google, 2011a). Eric Schmidt, Google's former chief executive officer, has explained that at Google.com there is information delivered that is legal in the United States, and illegal in other countries. Google asserts that when a result is on Google.com it is essentially controlled by Google U.S. and under jurisdiction of U.S. law. Google thus offers

local search engines that are compliant with local laws. One of the earliest and most commonly used examples by Google executives (and by the search engine industry more generally) is that pro-Nazi material is illegal in Germany (and France), and Google omits those websites in their local domain search engines, Google.de and Google.fr (Schmidt, 2009; Whetstone, 2010). Google also abides by national youth protection laws, for instance in Korea by enabling Safe Search by default. In such cases, Google's results page states the number of returns that have been removed for legal reasons (Whetstone, 2010). Google.cn is the most well-known as well as controversial instance of localization, whereby Google's Chinese engine filtered results drastically. It took a novel approach in 2010 by redirecting users of Google.cn (China) to Google.com.hk (Hong Kong), where Google does not filter, according to the company (Drummond, 2010a; Drummond, 2010b).

There is, of course, further literature to draw upon when studying national webs, from the pioneering ethnographic study of the national web of Trinidad and Tobago, where not global but rather Trini culture is performed, to well-known works on media as organizing national sentiment and community more generally (Higson, 1989; Anderson, 1991; Miller and Slater, 2000; Ginsburg et al., 2002). In policy studies, too, national webs, or portions of them, are increasingly "mapped" to inform debates about the extent to which the web, and especially the blogosphere, organizes voice (Kelly and Etling, 2008; Etling et al., 2010). Of interest is the related work that seeks to build tools to circumvent censorship so that voice is still heard (Glanz and Markoff, 2011; Roberts et al., 2011). In library science, national webs are routinely constructed by national libraries and other national archiving projects, which also have considered how to define such a web (Arvidson and Lettenström, 1998; Arms et al., 2001; Abiteboul et al., 2002; Koerbin, 2004). There are variously sized national web archives. Countries that have legal deposit legislation for web content as well as books (such as Denmark) tend to have notably larger web archives than the countries that do not (such as the Netherlands) (PADI, n.d.; Lasfargues et al., 2008).

Defining national websites, and the implications for national web capture

Archivists' definitions of national webs and national websites are of special interest in our undertaking. How do national libraries define national webs and websites? What may we learn from their definitional work? For example, the Royal Library of the Netherlands, following similar definitions of a national website from archiving projects in other European countries, defines a website as "Dutch" if it meets one or more of the following tests.

What is a "Dutch website"? It is a Dutch website, if it is:

a) Dutch language, and registered in the Netherlands;
b) Any language, and registered in the Netherlands;
c) Dutch language, registered outside the Netherlands; or
d) Any language, registered outside the Netherlands, with subject matter related to the Netherlands (Weltevrede, 2009).

The above scheme for what constitutes a national website, or at least one deemed relevant for a national archiving context, has consequences for their collection. Here in the first instance we would like to discuss how a definition affects the collection technique, automated or by hand. If one were to begin with sites from the national domain (.nl), those sites in Dutch (and ones in other languages) may be automatically detected with software, and in the collection procedure, one would remove from the list .be sites (from Belgium, where Dutch, or Flemish, is also spoken), unless they treated Dutch subject matters. (Dutch national web archive users likely would be surprised to come upon Belgian websites stored in it for whatever reason!) The Royal Library's could be described, however, as an editorial approach, for especially websites related to Dutch subject matters and websites in Dutch but registered outside of the Netherlands (outside of .nl) pose particular challenges to automation, and working at scale. As a research practice, one would not be able to automate the detection and capturing of those sites; one would more likely create a list of them, before routinely capturing them over time. In the national web characterization studies, reviewed by Baeza-Yates

in 2007 and discussed at the outset, the national domain (known as the country code top-level domain, or ccTLD) is the organizing entity. In practice, however, many countries (or nationals) use URLs outside of their national domains, such as .com, .net and .org. As we note below, for Iran, sites with the .ir ccTLD in fact may not be the preferred starting points for demarcating a national Iranian web. As a case in point, in our data the percentage of .ir sites that is blocked is very low, compared to .com's, for example. Thus .ir seems to have characteristics that differ from other sites authored and/or read by Iranians.

In order to describe the considerations an analyst may have when beginning to demarcate a national web, and at the same time to direct these thoughts to the specificity of Iran, we first surveyed a selection of Iranian bloggers about the "Iranian web," and particularly the very ideas of an Iranian website as well as a national web.

We are particularly interested in contrasting definitions of a national web that are "principled" with those based on device cultures. By principled we mean *a priori* definitions of what constitutes a national web and a national website, such as the archivist's above. By device cultures we mean the webs that are formed by collecting and analyzing user data, and outputting leading sites of a country and/or language. We mentioned above some of the consequences of demarcating a national web when national websites of interest to archiving are based on formalist properties of their content. It becomes difficult to make a collection at any scale.

In preliminary research about the very notion of an Iranian web, a small survey, undertaken by a New Media M.A. student at the University of Amsterdam, was made of Iranian bloggers using Google Reader (Gooder) in the student's Gooder network (n=141) (Zarrinbakhsh, 2011). A variety of definitions of a national web were put forward, and the respondents were asked to choose which definition was best suited. (They could choose multiple answers.) From the beginning, the question was met with suspicion, as the term itself was seen as a possible ruse by the Iranian government to create its own Internet, and further isolate the country and the people, as the student reported. In comments on the question, it was written that the Internet is a "free sphere" and ideas of a national web would "limit" such freedom.

The questions read as follows: What is an "Iranian website"? It is an Iranian website, if it is:

a) Only in the Persian language
b) In Persian and other languages (and dialects) spoken in Iran
c) Authored by Iranians
d) Related to Iranian issues
e) Accessed by Iranians
f) National domain (.ir)
g) Returned by Google

Note first the expansion of considerations for what would constitute a national web beyond what we have related so far, both in national domain characterization studies but also in the case of the constitution of the Dutch web by the Royal Library. In particular, sites accessed by Iranians and those returned by Google are newly added candidate constructs of an Iranian web, where the former treats the Iranian web like a traditional media consumption survey: Which sites are most visited? The last question about Google's relationship with the Iranian web is more ambiguous. Google could be equated with the web generally, as its entry point. Or, one could find the Iranian web with Google.

As a whole, twelve percent believed that only Persian websites could be considered national websites. Thirty-one percent checked the box for Persian websites and other languages and dialects spoken in Iran. Forty-five percent thought that when Iranians produce the content, it could be counted in the area of a national web. Twenty-nine percent were of the opinion that everything related to Iranian issues is in the area of the Iranian national web. Nineteen percent were of the opinion that the websites accessed by Iranians show their national web. It should be noted that some people were very much opposed to this choice; they mentioned that every website can be accessed by anyone, so this item seems to be ill conceived. Four percent of the total respondents chose websites with the Iranian domain (.ir), implying that national web studies relying on the domain only would prove unpopular. Nine percent thought that websites that appear in Google search results make up the (Iranian national) web.

Finally, in a follow-up question addressing the issue of any difference between writing from inside or outside the country, approximately one-third of the respondents seemed to agree with the communications scholar, Gholam Khiabany:

If Iranian blogs are defined in terms of language, this means omission of a large number of Iranian bloggers who write in other languages, most notably English, while including a number of bloggers from Afghanistan or Tajikistan who write in Persian. Focusing on Iranian bloggers writing inside the country also leads to excluding a large number of Iranian bloggers writing in Persian outside Iran (2007:565).

On the basis of these survey findings, and extending Khiabany's thought, the analyst concluded that a national web could be defined as one that is authored by Iranians, no matter their location or language in which they write, and no matter the subject matter. Such a definition of the national web appears to include sites with content authored by Iranians outside of Iran in languages other than Persian, on issues that may not be related to Iranian affairs. This is a case whereby the definition makes it nearly impossible to demarcate an Iranian web! In any case, detecting sites authored by Iranians outside of Iran in languages other than Persian would require manual work. It may be worth noting that the definition adhered to by the Royal Library of the Netherlands also required manual work, but did not expand its definition of Dutch sites to sites authored by Dutch people abroad in languages other than Dutch, unless the subject matter was Dutch-related.

Having considered and discussed what we have termed principled approaches to defining national websites and webs, we instead chose to analyze the outputs of devices, which we come to again in more detail shortly. That is, methodologically, we do not begin with *a priori* definitions of what constitutes an Iranian website, or the Iranian web, however fascinating in a formalistic and ontological sense. Rather, as we explain and eventually defend, we rely upon the URL recommendations made by dominant web devices and platforms, which through different algorithms and logics are deemed relevant for a specific country and/or language.

Our contribution to national web studies informs the literature on national web characterization, as discussed in the opening, as well as on policy studies (and political science) about the organization of voice online. It also contributes to media theory and web studies by putting forward the national web as object of study. The overall approach is not only conceptual but also empirical, in that we seek properties of na-

tional web spaces that are indicators of conditions on the ground. Such properties could be how responsive a national web is at any given time, and how accessible. Are responsive sites also fresh, or recently updated? Are sites that are blocked still responsive and fresh? We are also interested in more than the technical web data sets, and how they may be repurposed for social study. As alluded to already, for Iran in particular the content of websites is carefully monitored by the state; websites may be blocked and website authors may be pursued. In the following, we put forward an approach to demarcating a national web, in order to study its current conditions, including analysis of changing degrees of expression and voice (2009-2011).

Demarcating the Iranian web: Studying the outputs of device cultures

The purpose of the research is to demarcate a nominal Iranian web, and analyze its condition, thereby providing indications of the situation on the ground. By nominal web we mean one that is predicated on the means by which it is organized by online devices and platforms as well as retrieved, both by the user and by the analyst. Here we have chosen to demarcate an Iranian web through multiple, dominant online approaches for indexing and ordering that "go local," and privilege language, location and audience, broadly speaking. Working in July 2011, we found that the web given by three crowd-sourcing platforms aimed at an Iranian audience differs from that yielded by a marketing tool for Persian-language advertisers, a surfer pathway aggregator of users in Iran, and a search engine delivering .ir sites as well as other top-level domain sites from the "region," even though each purports in some general or specific sense to provide the Iranian web. Ultimately we have chosen to write about the Iranian webs in the multiple, and discuss each web's characteristics. We thereby addressed an issue faced by the analyst when formulating where to start collecting URLs, be it in terms of compiling seed URLs to crawl, stringing together keywords and operators to form a query, consulting lists of top blogs by inlink count, top URLs by rating or top websites by hit count, etc. For our analysis we selected the outputs of the well-known aggregators of Iranian or Persian-language websites, in a sense not choosing one starting point, but retaining them all—or at least a number of significant ones.

We also took decisions with respect to dealing with the idea that a sample of the Iranian web would follow (only) from knowledge of its population. As we discuss, in the national web research area one may be confronted by expectations of knowing the population of a web (in terms of the number of websites, and some categorization of their types), and being able to make a sample from it and from its types. In thinking through such an undertaking, one may port scan the Iranian IP ranges and establish whether IP addresses respond to the standard HTTP and HTTPS ports 80, 8080, or 443. One would count how many web servers are active within a specific IP range, and in a second step roughly estimate the number of domains. Alternatively, one may consider approaching the Iranian Internet authority or Iranian ISPs for their data. Or, one could crawl a seed list of URLs, or multiple lists, in snowball techniques, and subsequently sift the large catch by language-detection software and/or whois lookups. When one begins to rely on web services that have ceilings or have issues with spammers and scrapers (which is most if not all of them), then the challenges of (relatively) big online data become apparent. One is unable to run batch queries without permission from corporate research labs, Internet administrative bodies and others. Just when it is becoming interesting, the research focus turns to the administrative, legal and social engineering arenas, bringing everything to a standstill. Research that gains the access, and finishes the large collecting and sifting project, become great achievements in themselves. Whilst we have undertaken one medium-scale scraping and querying exercise for this research project, we largely avoid the techno-administrative arena we refer to above, and instead seek to make use of what is available to web users. We make a conscious choice in favor of relatively small data.

Furthermore, we would like to make a case for a method to demarcate a national web (or "webs") that is sensitive to the variety of ways one enters web space by belonging to particular device cultures, which we largely equate with engine and platform operations, instead of in an ethnographic sense (where an object may have a spirit, for example). Generally, we introduce national web demarcation methods that repurpose web devices that not only "go local," but also capture device cultures. In short, we are interested in capturing national device cultures. Repurposing web devices has two methodological advantages. First, popular devices may be viewed as mediating and quantifying specific usage. The devices do so by recursively soliciting user

participation in content production and evaluation. They calculate the most relevant websites by aggregating links, clicks, views and votes, thereby outputting collectively privileged sources. Second, the definition of an Iranian web is outsourced to the big data methodology used by devices to order content, which combines algorithmic techniques with large-scale user participation. Relatively small data sets are obtained from the output of these big data devices. Put differently, the repurposing of web devices is both a strategy for the small data researcher to sample from a big data set as well as a means to have samples that represent specific outlooks on how to organize and order web content, as we explain in our discussion of the privileging of hits, links, location, likes and other measures by the platforms and devices under study.

In the analyses, we wish to chart language and other formal features that are in each Iranian web. More conceptually, in our particular approach to national web studies, we also would like to discuss which portions of the web are healthy, in the sense of (still) online and active, and which are broken, in the sense of unresponsive. Additionally, we are interested in the extent to which each is censored or filtered by the state, and whether there is a relationship between responsive (and fresh) websites and filtered websites. In order to pursue the question of whether censorship kills content, which we have formulated in a previous (and preliminary) project on the Tunisian web (prior to the "Arab Spring" of 2011), we developed means to chart changes to a special part of the Iranian web over time. Here we use time-series data from Balatarin, a leading crowd-sourced platform which we scraped, comparing the significant URLs voted up around the presidential elections in 2009, with those of the same time period in 2010 and 2011. First we ran the hosts through proxies in Iran so as to check for indications of blocking. Generally we found that Balatarin's collection of URLs is particularly susceptible to blocking. We also analyzed the use of particular words ("fiery language"), in order to make findings about voice online in times of suppression and repression. We are particularly interested in the relationship between the use of that language on websites and the blocking of those same sites. Do the authors of the webpages continue to use language that would have their sites blocked? Generally, we discuss our findings in terms of the strength, clarity and volume of voice, which we describe. Prior to reporting on the longitudinal analyses, first, in the following, the indexing and ordering mechanisms of

the web platforms and devices relevant to the Iranian space are described. The data culled from these platforms and engines are employed to characterize the web types on offer.

Device cultures: How websites are valued and ranked

The early web was organized by amateur as well as professional link list makers, who took on the mantle of a librarian or specimen collector, and made directories of websites, organized by category. Professional or "pro-am" website categorization by topic remains, in the larger-scale directories such as Yahoo! as well as in smaller-scale collections, though the practice arguably has declined in the face of the other methods (which we describe here) that have become more and more settled as dominant approaches online for valuing websites (Deuze, 2007; Bruns, 2008). These approaches of valuing websites we couch in technical as well as politico-economic terms as the "hit economy," "link economy," "geoweb," "crowd-sourcing," and the "like economy," which highlights what is counted, by whom and/or where. Crowd-sourcing, a term coined by the Internet trade press that derives from the practice of outsourcing, also has been described as the "worker-bee economy," where both the so-called wisdom but also the labor of the crowd pollinates the beneficiary, often a Web 2.0 company or service (Howe, 2006; Moulier-Boutang, 2008). The other term we employ, the geoweb or locative web, has less of the connotation of a particular kind of economy, yet contains the means by which sites are sourced.

The hit economy, once exemplified by the hit counter on early websites, ranks sites by the number of hits or impressions, where unique visitors count. For such a view we have chosen DoubleClick Ad Planner by Google (referred to here as Google Ad Planner), which is a service that ranks sites by audience for the purposes of advertisers. Whilst "Iran" is not among the countries listed (which likely owes to a combination of the lack of a .ir local domain Google as well as the U.S. economic sanctions against Iran), Persian-speaking is among the site type categories in the available audience analytics. Thus one Iranian web would be comprised of those sites that reach a Persian-speaking audience, as collected and ranked by Google Ad Planner. Using the options available, 1,500 unique hosts for a Persian-speaking audience were collected from Google Ad Planner.

The "link economy" is a term that describes the rise of PageRank and other algorithms that value links (Rogers, 2002). It also captures a shift in URL ranking logics away from an advertiser's model (hit-counting) to a more bibliographic or scientometric manner of thinking (citation or link-counting). The term is used to characterize Google Web Search, however, much of the other main component to the algorithm apart from link-counting is user click-throughs. Searching Google for .ir sites (including .ir's second level domains) as well as Iranian sites in generic top-level domains in Google's regional search, yielded some 3,500 hosts.[3]

Alexa, like other companies offering browser toolbars, collects user location data such as a postal code upon registration, and once the toolbar is installed, tracks websites visited by the user (see Figure 1). It thereby keeps records of the sites most visited by user location. Alexa furnishes a list of the top 500 sites visited by users in Iran.

Crowd-sourced sites such as the most well-known (Balatarin) and its emulators (Donbaleh and Sabzlink) require registration before the user may suggest a link, which is then voted upon by other registered users. Those URLs with the most votes rise to the top. For this exercise, we collected approximately 1,100 different hosts from Balatarin, 2,850 from Donbaleh and 2,750 from Sabzlink.[4] In the following analyses we grouped the two crowd-sourcing platforms Donbaleh and Sabzlink, for they share the device culture (crowd-sourcing). Together they resulted in 4,579 unique hosts. We treated the other crowd-sourcing platform, Balatarin, separately. The special treatment arises from its status as a highly significant Iranian website.

Launched in 2006, Balatarin is considered the first Web 2.0 site in Persian, and has been recognized as one of the most popular Persian websites in 2007 and 2008 (Wikipedia, 2011). It also has been pivotal for the Green Movement in the opposition before and after the Iranian presidential elections in 2009 (Iran Media Program, 2010). The recognition of Balatarin as a platform for the opposition also provides the opportunity to employ it as a barometer in studying the continuing strength, clarity and volume of that voice. Do the websites that are recommended on Balatarin continue to express themselves critically, or have they discontinued the use of language critical of the regime? By strength of voice, we examine the continued use of the words. To study clarity, we examine whether the words they choose are fiery and side-taking or coded. Volume is whether there are more and more voices that are using the words: Is the chorus (so to speak) growing louder?

The introduction of the like button and other social counters in social media has brought with it what one may term the "like economy," which values content based on social button activity (Gerlitz and Helmond, 2011). Likekhor, as the name suggests, ranks websites by likes; the likes are tallied from Google Reader users who have registered with Likekhor. Google Reader, or Gooder (as some Iranian users call it), is of particular interest because through it one has been able to read the contents of websites that are otherwise filtered by the state. Google Reader thus effectively acts as a proxy to access filtered websites. At Likekhor the focus is on blogs, pointing up a relationship between Google Reader users and bloggers, or blog readers. From Likekhor we extracted a list of 2,600 hosts, which are collected from a page where all blogs on Likekhor are listed.

Figure 1: Alexa toolbar installation and registration process, with field for user's postal code, August 2011.

---

3 Google.com's web search was chosen for its dominance in Iran among users of search engingooes. Data from 2010 list search engine market shares in Iran as follows: Google 90.78%, Yahoo 4.97%, Bing 3.64%, Ask Jeeves 0.46%, AOL 0.07% (MVF Global, 2010). Another marketing research firm lists 2011 market shares in Iran as Google 87.15%, Yahoo 7.27%, Bing 4.16%, Ask 0.70%, AOL 0.12% and Lycos 0.01% (Net Applications, 2011). According to Alexa in October 2011, Google.com is the most visited site in Iran, followed by Yahoo.com. We employed site queries in Google.com for the top level (site:.ir) as well as the second level domains (e.g., site:.co.ir), and concatenated the results. The query technique did not allow for the re-directing to a local domain oogle. Because cookies had not been retained, it also did not allow for the personalization of the results.

4 In order to compare the different platforms we chose to compare hosts instead of full URLs. That is, for Balatarin, we harvested all the URLs listed on the 150 pages of "hot" links, resulting in 1102 unique hosts.

Thus, in July 2011, we collected more than 10,000 unique hosts through platforms and devices significant to Iranian users (Google Reader, Google Web Search and the crowd-sourcing platforms) and two that provide ranked lists of Iranian or Persian-speaking sites (Alexa and Google Ad Planner) on the basis of data collected from users located in Iran (Alexa) or from Persian-writing users (Google Ad Planner). We will characterize these Iranian webs individually as well as collectively. We have chosen not to triangulate them, for very few websites recur across them.

Analyzing the characteristics of the Iranian web: Language and responsiveness

One area of research that we build upon is web characterization studies, where one of the main difficulties repeatedly discussed is how to obtain a representative sample of a national web or other web types. According to Baeza-Yates and colleagues, the three common types of sampling techniques used in web characterization studies are "complete crawls of a single web site, random samples from the whole web, and large samples from specific communities" (2007: 1). For national webs, which the authors consider to be specific communities, the list is comprised of websites with the same ccTLD (country code top-level domain). For many national webs, however, such delimiting would be too partial, certainly for countries where generic top-level domain usage is prevalent. Our approach seeks to retain the .com's, .org's, .net's, etc. when deemed relevant for Iranians and Persian-speakers by the devices and platforms upon which we rely.

To the sampling techniques described above, we thus would like to add a fourth type which could be called multiple aggregator site scraping, or more conceptual-ly, device cultures. Google Ad Planner, Alexa, Google Web Search, Likekhor (Google Reader) as well as the crowd-sourcing platforms (Donbaleh, Sabzlink and Balatarin) make available (either through query results or dynamically-generated listings) websites that are relevant for Iranians and Persian speakers. In our case, with the exception of the searcher's web (gained through .ir and generic TLD queries in Google's region search), the percentages of .ir sites among the significant hosts outputted by the devices are relatively low (see Table 1). The crowd-sourced web references the fewest .ir sites at just over 10 percent, whilst that of both the advertiser's as well as the geoweb, or web of surfers in Iran, has the highest percentage at about 25 percent. As noted earlier, the .ir sites in our overall collection of URLs are much less likely to be blocked than the .com sites. Of the websites that are tested and found blocked from inside Iran, 80 percent are .com, followed by .net with 6 percent and .org with 4 percent. The ccTLD .ir has 3 percent of all censored hosts.

Having reviewed how samples are generally made, Baeza-Yates and colleagues compared the ten national web studies, in order to arrive at a core set of measures that are shared across many of them (see Table 2). Our characterization of the Iranian web (or webs) has a particular point of departure that benefits from the metrics on offer. In reference to the metrics in Table 2, in the category of content our project shares interest in language, page age and domain analysis (albeit top-level), and in the category of technology, relies on HTTP response codes. The codes yield what we refer to as "responsiveness," which we consider a basic health metric, together with page age, the freshness measure. There are other metrics that we do not employ, though we would like to mention how to do so. Brokenness could be gleaned from link validators, where it would refer to broken links on a site. Additionally, establish-

Table 1: Percentage of .ir sites in top websites collected from device cultures relevant to Iranians and Persian-speakers, July 2011.

Percentage	Iranian web	Absolute numbers
25%	Alexa (Geoweb)	126 of 496 hosts
24%	Google Ad Planner (Advertiser's)	370 of 1,525 hosts
16%	Likekhor (Blogger's)	397 of 2,541 hosts
12%	Donbaleh/Sabzlink (Crowd-sourced)	535 of 4,579 hosts
11%	Balatarin (Crowd-sourced)	116 of 1,102 hosts

Table 2: Metrics commonly used in national web characterization studies according to Baeza-Yates et al., 2007. Bold indicates metrics used in this study, but we analyze the top-level domain rather than the second-level domain.

Content	Link	Technology
Language	Degree	URL length
Page size	Ranking	**HTTP response code**
Page age	Web structure	Media and document formats
Pages per site		Image formats
Sites & pages per domain		Sites that cannot be crawled correctly
Second-level domain		Web server software
		Programming languages for dynamic pages

ing whether websites are "parked" or "hacked" may serve as measures of abandonment by previous owners. Compared against proxy data, parked or abandoned site analysis may be used to make claims about the effectiveness of censorship, or suppression of voice. Fitness could refer to the "validity" of code, or correct implementation; Baeza-Yates and colleagues refer to site structure, and its "correctness" for a crawler. Other metric types more in the realm of political economy that are of interest to us in expanded undertakings are available. For example, media, document and image formats could give us an indication of the extent to which a national web is proprietary, which from certain perspectives is a health issue.

The Iranian web and its languages

One basic metric seeks to measure the composition of languages in the Iranian web (see Figure 2). Persian is of course the official language in Iran; the Unicode system incorporated Persian script in 2001, and it can be detected (Amir-Ebrahimi, 2008). For language detection of websites we built a custom tool that makes use of alchemyAPI's language detection functionality, and is able to detect Persian as well as the other languages, though not all languages spoken in Iran, as we relate.[5] In a second step, the results are manually checked.[6] Two out of three sites in the Iranian web,

in total, are in Persian; English is second with one of five. Of interest are the proportions of Persian used in the various webs. The results show that the blogger's space, Likekhor, is on top with 91 percent of the sources in Persian, followed by Alexa's Iran-based surfer's web with 83 percent and the crowd-sourced web with 73 percent. At the bottom are the advertiser's web with 62 percent, and Google Web Search with 52 percent. Balatarin, the special case, has 75 percent in Persian. Thus there is significant difference between the webs, including, notably, a Persian-dominant blogosphere (if the Likekhor list may serve as a short-hand reference to such).[7]

Here we can begin to discuss the kinds of webs that one were to capture and analyze if one were to define the Iranian web or an Iranian website *a priori*, and seek it according to a formal definition, a subject matter raised earlier with respect to the web archivist's formal conditions of a national website (in the Dutch example) as well as the survey respondents' ideas of a national web (for Iran). The blogosphere and to a slightly lesser extent the geoweb (based on surfers in Iran) are most closely related to ideas of an Iranian web as Persian-speaking only, though in that case between them there still would be an average of more than 10 percent of the non-Persian websites to be reckoned with. The Iranian webs with larger percentages of non-Persian sites are the advertiser's as well as the regional web

---

5 The language auto-detection functionality is provided by alchemyAPI, which for academic researchers allows 30,000 queries per day. The tool is at http://www.alchemyapi.com/api/lang

6 We manually checked the results which returned sites as English or unknown, and corrected any errors. We have not explored further why dual-language sites are considered as one particular language by alchemyAPI. We also would consider using Google Translate as a language detector. The

'unknown' tags in the cloud indicate that neither the language detection tool nor the researcher was able to determine the language, for in most cases the site was no longer online.

7 Additionally, the Iranian webs show various degrees of language distribution, with Alexa being the least diverse with six languages and Google Web Search the most with 36 languages.
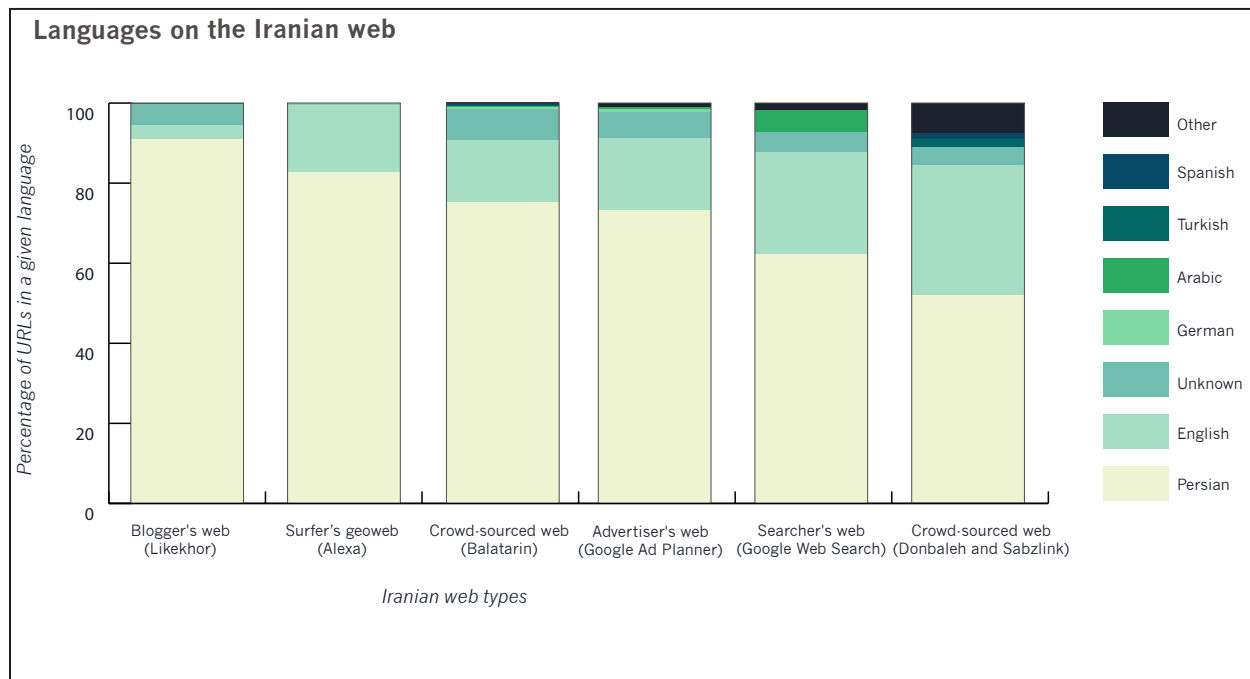
Figure 2: The distribution of languages on the Iranian web, August 2011. Graphic by the Digital Methods Initiative, Amsterdam.

(from Google's advanced search region option). The advertiser's is the web accessed by Persian speakers as detected by the signals Google compiles on its users and the content it indexes (Google Ad Planner). Both have far higher percentages of non-Persian sites, especially English, though we did not attempt to investigate whether these sites are authored by Iranians, or concern Iranian affairs, however that may be defined.

There is another web one could conceive of *a priori*, which also would have implications for the method by which one would construct the object of study. Being all-inclusive in terms of the languages spoken in Iran (Armenian, Assyrian Neo-Aramaic, Azeri, Kurdish, Lori, Balochi, Gilaki, Mazandarani, Arabic and Turkmen) has consequences for the capturing techniques; of the secondary languages spoken in Iran, the language detection tool employed in this study detects only Armenian, Arabic and Azeri, and not Assyrian Neo-Aramaic, Kurdish, Lori, Balochi, Gilaki, Mazandarani or Turkmen. To compile such sites, one would rely on specialists' link lists, though we did not pursue the matter any further.

The Iranian web and responsiveness

To analyze the responsiveness of the Iranian webs we retrieved the HTTP response status codes (of the some 10,000 unique hosts) from the Netherlands with a custom-built tool. The inputs to the tool are the lists of hosts per web that previously were collected. Analyzing the results returned by the response code tool, we found that there are eight commonly returned codes in the Iranian web spaces (see Figure 3).[8] The 400 class of status codes indicates that the client has erred; "404 not found" (which means that the content is no longer available) is considered the strongest indication of unresponsiveness. "400 bad request" means that there was an error in the syntax, and "403 forbidden" indicates that the server is refusing to respond. Commonly returned response codes besides the "200 OK" status (standard successful HTTP response) are two redirecting response codes: "301 moved permanently" and "302 found." Redirecting is not necessarily an indication of unresponsiveness, and can have a range of reasons, including forwarding multiple domain names to the same location, redirecting short aliases to longer

---

8 The http status codes are explained on the dedicated Wikipedia entry, http://en.wikipedia.org/wiki/List_of_HTTP_status_codes (accessed 14 July 2011).

The health of the Iranian web

Advertiser's web
(Google Ad Planner)

Blogger's web
(Likekhor)

Surfer's geoweb
(Alexa)

Searcher's web
(Google Web Search)

Crowd-sourced web
(Balatarin)

Crowd-sourced web
(Donbaleh and Sabzlink)

Legend:
- Other
- 502 Bad Gateway
- 500 Internal Server Error
- 410 Gone
- 404 Not Found
- 403 Forbidden
- 401 Unauthorized
- 400 Bad Request
- 0 Connection Problem
- 200 OK
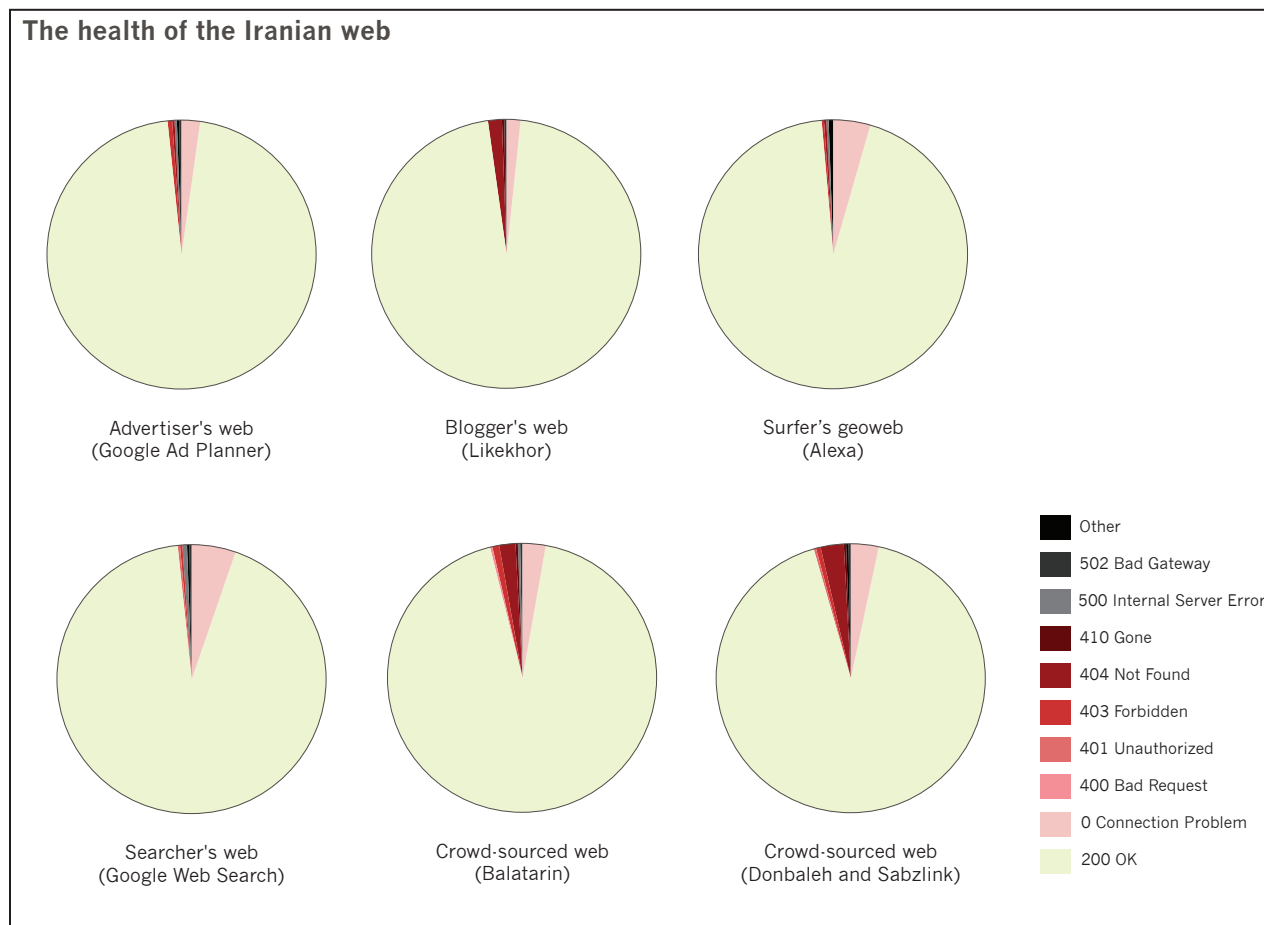
Figure 3: The health of the Iranian webs measured by HTTP response codes in the Netherlands, August 2011. Graphic by the Digital Methods Initiative, Amsterdam.

URLs, or moving a site to a new domain.[9] It also may be an indication of a parked website. However, redirects also may be "soft 404" messages to hide broken links (Yossef et al., 2004). In the current study, both 301 and 302 were followed if a location header was returned, which mostly resolved in 200 and 404 response codes. "0 connection problem" indicates that the tool was unable to connect to the server; the server may no longer exist, or it may mean that that the site did not respond within 60 seconds.

The findings of this portion of the study indicate, first, that the Iranian webs are relatively healthy overall. The crowd-sourcing webs of Donbaleh/Sabzlink and Balatarin have 92 and 94 percent of the sites resolving, respectively. The advertiser's space, followed by the blogger's space, delivered by Google Reader users, have the cleanest bills of health, with 96 and 95

percent of the websites resolving. Thus the vibrancy of the (Persian-language) advertiser's space and the blogosphere as well as the crowd-sourced webs is a finding.

The Iranian web and Internet censorship

Arguably, web devices are among the most well-informed censorship monitoring instruments. Search engines and platforms receive requests for deleting content—either specific URLs, specific queries or more general instructions—thereby inviting the creation of an ongoing blacklist as well as a censorship index. For example, it has been reported that to adhere to Chinese government censorship instructions (prior to the redirect to .com.hk), Google engineers "set up a computer inside China and programmed it to try to access websites outside the country, one after another. If a site was blocked by the firewall, it meant the government regarded it as illicit so it became part of Google's

9 URL redirection is explained on the dedicated Wikipedia entry, http://en.wikipedia.org/wiki/URL_redirection (accessed 14 July 2011).
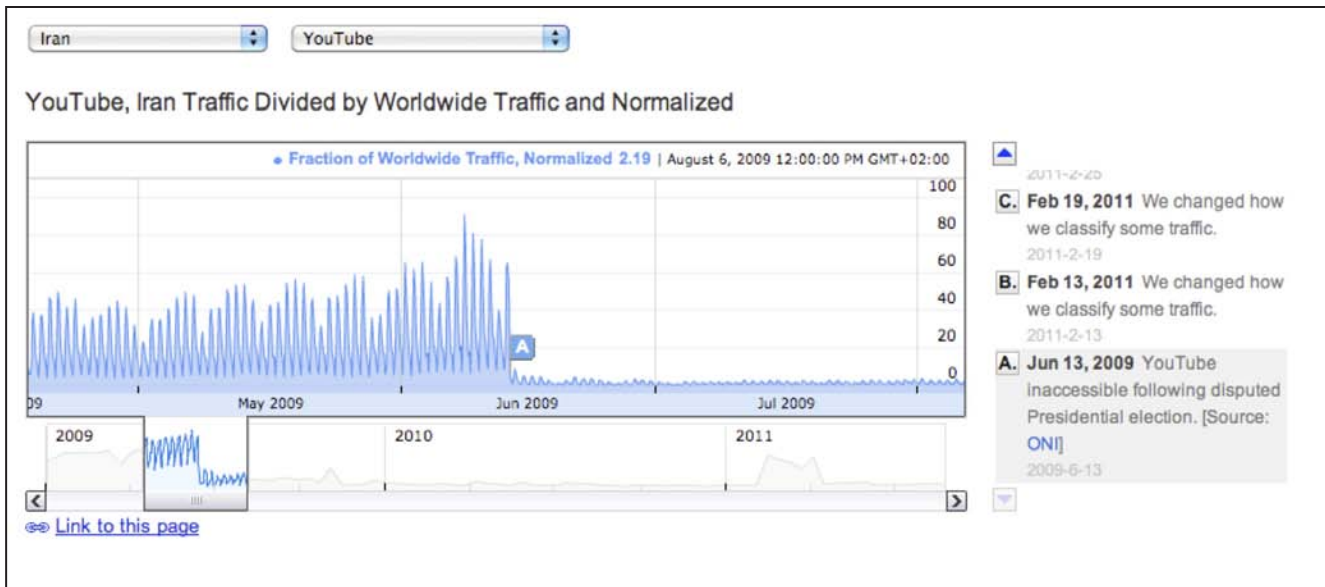
Figure 4: Iranian traffic to YouTube comes to a standstill after the 2009 presidential elections. Source: Google, 2011b.

blacklist" (Thompson, 2006). In the case of the Iranian web, which is among the most aggressively censored webs in the world, there are no reported requests for removal from the government (Open Net Initiative, 2009; Google, 2011b). The graph in Figure 4, however, shows how Iranian traffic to YouTube increased in the run-up to the presidential elections in June 2009, before coming to an almost complete standstill one day after. The question of interest in this study is to what extent blocking important sites has had on the health of the Iranian webs. In the following, the Iranian webs collected are checked for availability inside Iran by using proxies. Subsequently, these findings are compared against the basic health measures, responsiveness and freshness. As mentioned above, one of the more remarkable findings is that a large portion of the Iranian blogs is blocked, yet continues to respond, and is fresh.

The Censorship Explorer tool, which we have made available at http://tools.digitalmethods.net/beta/proxies/, lists (fresh) proxies by country, and may be used to check for censored websites. The tool returns website response codes, or loads the actual websites in the browser, as if you were in the chosen country in question. As a starting point in the censorship research procedure, one often checks website responsiveness in a country that is not known to censor (Iranian) websites (in this case, the Netherlands). Subsequently, one runs lists of hosts through proxies in Iran, and logs the response codes. If the response code is 403 forbidden, while the response code is 200 OK when connected

from the Netherlands, it is understood as a strong indication that a site is blocked.[10] Although testing via proxies does not guarantee a replication of the average user experience, response code checks through proxies give indications of specific types of Internet censorship, i.e., URL and IP blocking through techniques such as TCP/IP header filtering, TCP/IP content filtering and HTTP proxy filtering (Murdoch and Anderson, 2008). (There are other known filtering techniques that are more accurately detected by other means, including DNS tampering and partial content filtering.) Often multiple proxies are used, allowing the researcher to triangulate proxy results and increase the trustworthiness of the results. For example, "0 Connection Problem" may be a proxy problem, but may just as well be that the censors return an RST (reset) package, which resets the connection, effectively dropping it (Villeneuve, 2006).

---

10 Typically, when an Iranian proxy returns 403 forbidden for a particular site, one is presented with an iframe loading http://10.10.34.34/?type=Invalid%20 Site&policy=MainPolicy which 30 seconds later redirects to http://peyvandha.ir, the site run by Ministry of Culture and Islamic Guidance. The former is only accessible from within Iran, and the latter contains a directory of recommended or approved sites, a list of reasons for banning a site, and a form to report a website thought to be in violation of the Iran's computer crimes law. While in this study we used response codes as strong indicators of blocked sites, we also conducted additional tests concerning the relationship between the 403 forbidden response and the presence of the block page URL. For blocked sites common to at least three lists (our test sample), we found that a 403 would be accompanied by a block page. It also may be noted that that http://peyvandha.ir ranks in the top 5 of Alexa's (surfer's) geoweb. As blocked sites redirect there, the site's high ranking provides a relative measure of the amount of traffic to blocked sites from within Iran.

Comparing multiple proxies can aid in confirming it is not a proxy problem. We used 12 proxies, which are hosted in six different cities in Iran and operated by a variety of owners, including Sharif University of Technology and the popular Internet service provider Pars Online. Concern has been voiced that it is "false to consider internet filtering as an homogeneous phenomenon across a country," considering that both the implemen Typically, when an Iranian proxy returns 403 forbidden for a particular site, one is presented with an iframe loading http://10.10.34.34/?type=Invalid%20 Site&policy=MainPolicy which 30 seconds later redirects to http://peyvandha.ir, the site run by Ministry of Culture and Islamic Guidance. The former is only accessible from within Iran, and the latter contains a directory of recommended or approved sites, a list of reasons for banning a site, and a form to report a website thought to be in violation of the Iran's computer crimes law. While in this study we used response codes as strong indicators of blocked sites, we also conducted additional tests concerning the relationship between the 403 forbidden response and the presence of the block page URL. For blocked sites common to at least three lists (our test sample), we found that a 403 would be accompanied by a block page. It also may be noted that that http://peyvandha.ir ranks in the top 5 of Alexa's (surfer's) geoweb. As blocked sites redirect there, the site's high ranking provides a relative measure of the amount of traffic to blocked sites from within Iran. tation and user experience of censorship may vary by city, ISP, or even by computer (Wright et al., 2011: 5). Taking note of this concern, we selected proxies from different cities and ISPs, and subsequently considered the response code returned by the majority.

The proxies used for this research:

217.219.115.133:80	ITC, Tehran, Esfahan
91.98.137.196:80	Sharif University of Technology, Sharif, Khuzestan
78.39.55.11:3128	ITC, Fars, Shiraz
91.98.137.196:3128	Pars Online, Tehran, Esfahan
80.191.120.129:3128	ITC, Tehran
213.217.43.82:8080	Pars Online, Pars, Tehran
217.219.115.137:80	ITC, Tehran, Esfahan
217.219.97.11:3128	ITC, Shiraz, Fars
80.191.122.11:3128	ITC, Shiraz, Fars
80.191.227.243:3128	ITC, Ahwaz, Khuzestan
188.136.241.2:3128	Ariana Gostar Spadana, Esfahan,
188.136.156.116:3128	Ariana Gostar Spadana, Gostar, Hamadan

The results show that approximately 5 percent of the searcher's web (179 out of 3547), 6 percent of the geoweb (29 out of 496) and 16 percent of the advertiser's web (238 out of 1,525 hosts) are blocked. The crowd-sourced web has just over 50 percent of the web blocked, with 2,382 of 4,579 hosts. Balatarin is the most aggressively censored Iranian web space with 57 percent blocked, or 623 of 1,102 hosts, followed by the other two crowd-sourcing platforms—Donbaleh and Sabzlink—with more than half of the hosts blocked. Google Reader's web, which in the research work thus far is standing in for the Iranian blogosphere, has 1,127 of 2,541 sites (44 percent) returning the "403 forbidden" code (see Figure 5).

As discussed above, the blogger's web is largely Persian language, and is one of the most responsive of all the webs under study, with 95 percent of the sites returning 200 OK response codes. Moreover, it speaks for the use of Google Reader usage as a vibrant censorship circumvention culture. This study appears to render visible censorship circumvention at a large scale, or at least blocked websites are still online. Of the webs checked for filtering, the crowd-sourced sites as well as the Likekhor listing are the most blocked, raising the question not only of the substance of those spaces (we treat Balatarin's below), but also the convenience of the platforms as URL lists for monitoring. Whilst many sites are blocked, and still responsive, we are interested in examining those blocked sites for other signs of health: Are they fresh? If the sites are blocked, yet responsive and fresh, we have a strong indication of the ineffectiveness of censorship (to date).

The Iranian web and freshness

Having identified the spaces of particular interest to us (crowd-sourced as well as the blogger's webs), and finding that they are highly responsive as well as heavily blocked, we are interested in pursuing further the question of whether censorship kills content. Or, despite having their sites censored, do the bloggers keep on blogging, and does the crowd keep posting, and rating? Is there an expectation that the readers can routinely circumvent censorship, and thus the content can continue to be recommended, commented on, etc.? Apart from the responsiveness test (which found nearly all of the websites online), we would like to know whether they are active. Is the content on the websites fresh? We are studying a subset of the webs—the blocked sites in the crowd-sourced and the blogger's webs. To determine how fresh these sites are, for each host (per

list) we ask the Google feed API (application programming interface) whether each site has a feed (e.g., RSS or atom). If it does, we parse the feed with the Python Universal Feed Parser library and extract the date of the latest post.[11] Overall, 63 percent (5,147 of the 8,222) of the three webs have feeds. Of the blocked sites in these webs, 71 percent (2,986 of the 4,189) has a feed. For Balatarin, the percentage of blocked sites with a feed is 79 percent (504 of 639 blocked hosts), for Donbaleh/ Sabzlink 68 percent (1,630 of 2,413) and for Likekhor 75 percent (852 of 1,137). These are the sites to be checked for freshness.

What constitutes a fresh site? We turned to blog search engines for advice about staleness. In an FAQ about blog quality guidelines, Technorati states that they "only index 30 days' content, so anything older than that will not appear on Technorati" (Technorati, 2011). Similarly, search engine and analytics system for blogs—Blogpulse—takes 30 days as a measure of fresh content: "A blog's rank is based on a moving average of its citation counts over the past 30 days" (Blogpulse, 2011). Thus, freshness is here considered as having at least one post published via a feed in the last month, counted from the moment we last checked for blockage. Would we expect these sites to be fresh? To draw our findings into stark relief, it is of interest to note that the well-known survey conducted by Technorati in 2008 found that only about 7 million of the

133 million blogs it follows had been updated in the past four months. The *New York Times* wrote that the finding implied that "95 percent of blogs [were] essentially abandoned, left to lie fallow on the Web, where they become public remnants of a dream—or at least an ambition—unfulfilled" (Quenqua, 2009). In this event we have found that 65 percent of the sites overall are fresh. In the crowd-sourcing platform Balatarin, 78 percent of the blocked hosts that have a feed (395 of 504 hosts) are fresh; in the crowd-sourcing web organized by Donbaleh and Sabzlink, 56 percent of the blocked hosts with a feed (915 of 1,630 hosts) are fresh. For the Likekhor list, 61 percent—or 525 hosts—have a post date of a month before they were tested and found blocked. The results confirm that there is hardly a general indication that censorship kills content on the Iranian web under study. On the contrary, the most severely censored Iranian webs are both responsive and rather fresh.

The Iranian web: Voice and expression

A substantive portion of the research project, touched upon in the introduction, concerns employing the web in order to gain indications of conditions on the ground. Indeed, it is another "health check" in the sense that we are interested in the strength of voice, and degrees of expression in hard times. Has voice been suppressed and expression become more dulled online over the past few years? How would one make a measure of such? This particular piece of research builds upon the work

---

11 The Universal Feed Parser downloads structured data feeds of many kinds, including RSS, Atom and CDF. It extracts post attributes, such as title, author, description, timestamp and link.
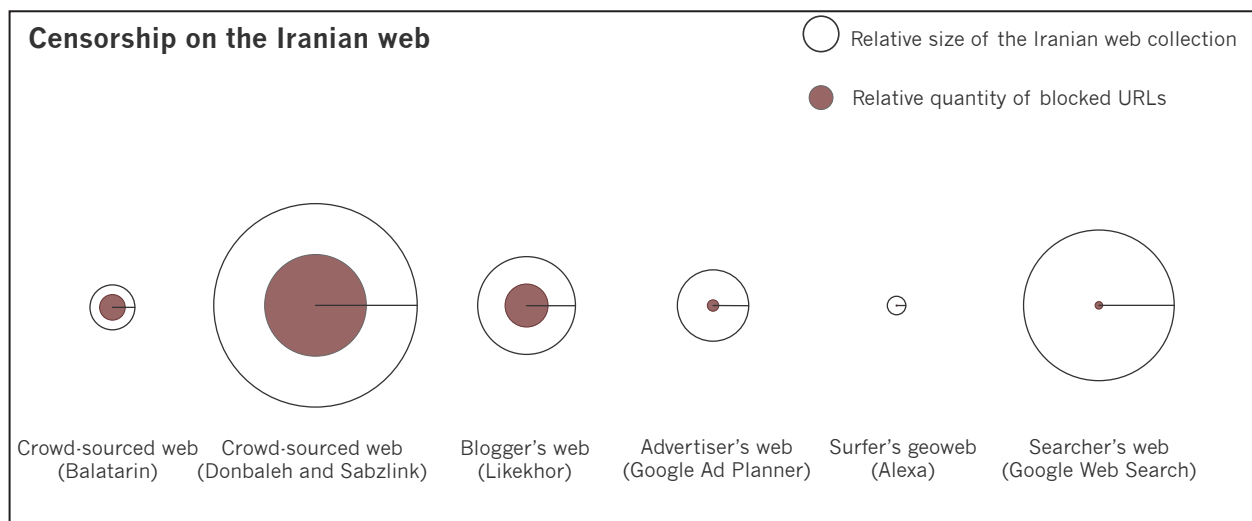

Figure 5: Censorship on the Iranian web, as measured through the share of 403 forbidden HTTP response codes. Data collected by the Censorship Explorer tool, Digital Methods Initiative (DMI), Amsterdam. Graphic by DMI.

on the Iranian blogosphere by John Kelly and Bruce Etling (2008). Prior to the 2009 elections, and the "color uprising" known as the green movement, they wrote about the repression in Iran, and argued that the Iranian blogosphere organizes voice in a particular way:

> Given the repressive media environment in Iran today, blogs represent the most open public communications platform for political discourse. The peer-to-peer architecture of the blogosphere is more resistant to capture or control by the state than the older, hub and spoke architecture of the mass media model, and if Yochai Benkler's theory about the networked public sphere is correct in relation to blogs, then the most salient political and social issues for Iranians will find expression and some manner of synthesis in the Iranian blogosphere. Future research could address whether or not this is true (Kelly and Etling, 2008:24).

We would like to inquire into "expression" by employing data from arguably the most significant Iranian website of the past four years, Balatarin. Balatarin, as discussed above, is considered here to be a set of URLs collected through a particular device culture. One of its salient features is the organization of the database that has been built up over time. Amongst other data held, Balatarin has the date that each URL was posted on its site since 2006. We scraped Balatarin's database in order to obtain the top URLs (from all topic categories) that appeared on the crowd-sourced platform, and the dates of their appearances. Subsequently, we downloaded the pages that were linked to from within the Balatarin posts, so as to be able to query them for a series of words, effectively making the work desktop research (searching for words). Our word list is comprised of what in Persian are called "smelly" words, or language that would be considered critical and out of order these days (see appendix).[12] We have devised

a scheme of term types that we thought would allow us to judge the effects of the suppression over time on voice and expression. We compiled 539 words, which included terms and phrases, as well as names of individuals. For the analysis, we used 235 of the 539, leaving aside phrases as well as many individuals' names, with certain exceptions such as Neda and Mousavi (see Figure 6). The list was sub-divided into three categories (a word can fall into multiple categories): fiery, side-taking, and coded. By fiery language, we mean language that would be (nearly intentionally) incendiary. If used, it would lead to the censoring of a blog or website. Side-taking language refers to terms that show (obvious) affiliation or alignment. The analysis of side-taking language enables not only an indication of the increasing partisanship of Balatarin (and the URLs its users recommend), but also to gain a sense of which language continues to be expressed, and which not, also as more and more websites are blocked by the state. Has that situation changed in the sense that more care is now taken in word choice? By coded or unspoken language, we specifically focus attention on language that is employed so as to not be blocked, or raise ire.

All of the words on the three lists have been chosen for their significance as forms of expression regarding some of "the most salient political and social issues for Iranians," as Kelly and Etling phrased it (2008:24). Our differentiation between types of words (fiery, side-taking and coded) was made so as to gain a sense of behavioral changes, such as the rise of coded language together with the decline of the use of fiery words, to give one example. Also, would oppositional voices grow weary, or move underground (and use fewer side-taking words)? Would the use of coded words become more prevalent as censorship (and harsher) activities expand?

We phrase our study as one concerning the organization of voice. In particular, we are interested in what we term the strength, clarity and volume of that voice, which we described above as continued usage of words over time, the choice of fiery and side-taking words over coded ones, and the sheer numbers of websites containing the words, respectively. Generally speaking, we found

---

12 The Persian term for "smelly" language referred to here is بودار. The word list was created through a collaboration by nearly 20 Iranian bloggers, whose blogs have been blocked by the state over the past three years. When their blogs were blocked, they began to compile a list of "smelly" words, such as open letter, manifesto, opposition party and political prisoner. To check the sensitivity (or, in our terms, fieriness) of the words, they would query each in google.com (http://google.com/search?q=smelly_word). If google.com was not blocked, and the query result was, then the term was considered censored (and indeed sensitive). Note that a blocked query result containing the fiery key word was not a criterion for inclusion on the word list, but rather an indication employed by the bloggers. It should be noted too that the words on the list are generally politically sensitive terms rather than routinely blacklisted key words related to alcohol, sex, etc., however

much the latter have been the object of study in larger inquiries into Internet censorship in Iran (Open Net Initiative, 2005; 2009). The choice of politically sensitive terms fits with our aim, which is not so much the general study of Internet censorship in Iran but rather the robustness of Iranian online expression of salient political and social issues.

that the use of the malodorous words has not declined, but rather has held steady, and actually increased, over the past three summers (2009-2011). As with fiery language, the use of side-taking language grew in volume over the years. Instead of self-censorship (of the fiery language), a greater use of coded words over time, or the quieting of side-taking, we found louder and louder voices, using all word types more and more frequently. The finding is all the more remarkable for the fact that there has been a concomitant rise in the blocking of the sites where the language is published. As sites are blocked, they are not dulled, but rather enlivened.

That is, an exploration of the data shows not a decline but a general rise in the use of all of our language types over time, comparing their occurrence in the websites posted on Balatarin between June and July 2009, and in those same months in 2010 and 2011. (The words are held constant; we generally do not add new smelly words as they become *en vogue*.) Focusing on the summer of 2009, around the date of the elections, there is, as expected, a significant rise in the use of fiery and side-taking (as well as coded) language after the elections on June 12, 2009. However, instead of a decline in subsequent years, as energies may flag and suppression spread, there is, as noted, only a rise in usage. The use of words termed fiery in the websites linked from Balatarin rose from 139,781 in June and July 2009 to 167,735 in June and July 2010 to 252,986 in June and July 2011. There is not only an absolute but a relative increase. No general chilling effect was observed for the other critical language used on websites that rose to the top on Balatarin. The use of side-taking language increased from 365,602 occurrences in June and July

Fiery Language 2011

Statistics for 2011-06-01 - 2011-08-01
139 distinct words recognized
249546 = the sum of word occurences (each word is only counted once per post)
89948 non-deleted and scraped posts
51412 posts with at least one *fiery* word
54169 posts with a blocked host
34252 posts with a blocked host and at least one *fiery* word

Color depicts the percentage of blocked hosts containing a particular word
0% - 20%
20% - 40%
40% - 60%
60% - 80%
80% - 100%

The tag cloud represents the *percentage* of posted links a particular word appears in.

[Persian tag cloud]

human rights (18.48) Mousavi (17.18) Karubi (16.72) Balatarin (15.89) political prisoners (11.42) strike (10.85) statement (10.71) rape (10.68) demonstration (10.4) download (10.24) Evin (10.02) struggle (10.01) protesters (7.71) dictator (7.63) march (7.27) mujahedin (5.9) torture (5.75) intelligence minister (5.27) homeland (4.58) hunger strike (4.55) coup (4.43) the green path of hope (4.26) Kahrizak (3.55) fraud (3.28) alliance (3.22) section 350 of Evin Prison (2.93) genocide (2.9) down with (2.84) Bahai (2.71) Radio Farda (2.64) BBC (2.47) collapse (2.08) Khordad 22nd (2.05) mercenaries (2.01) Azadegi (1.98) VOA (1.88) women's rights (1.77) totalitarian (1.64) civil society (1.36) political activists (1.24) HRANA (1.22) open letter (1.09) secular (1.03) People's Mujahedin of Iran (MKO) (0.98) Sepah's intelligence service (0.9) the brave sheikh (0.86) Tehran universitary dormitary (0.82) liberal (0.76) 80's (0.7) protesters (0.68) Taj Zadeh (0.67) solitary (0.65) green leaders (0.64) LGBT (0.63) Khavaran (0.59) wine (0.57) Neda Agha Soltan (0.57) Sane' (0.56) suffocations (0.55) proxy - anti filter (0.51) writing slogan (0.5) political activity (0.49) absolute ruler (0.44) Zeidabadi (0.43) Rahana (0.41) Risheha (0.38) Khordad 2nd (0.38) Aushura 88 (0.37) manifesto (0.37) civil disobedience (0.34) A.N (0.34) interrogator (0.33) liquor (0.31) Sane' Zhaleh (0.29) the green way council (0.27) Vakil Abad Prison (0.26) atheist (0.22) coup providers (0.22) opposition (0.2) Seyed Zia Nabavi (0.2) section 209 of Evin Prison (0.2) juice drinker (0.19) Kh.R [G.R] (0.16) mass destruction (0.16) Nedas (0.14) writing on paper money (writing slogans on paper money) (0.14) people demonstration (0.14) totalitarian (0.13) Labor Day (0.12) color revolution OR velvet revolution (0.12) open minded (0.11) say hello to the end dictator (0.11) nightly Allah O Akbar (0.11) Gene Sharp (0.11) Gharchak Prison (0.1) Where is my vote? (0.09) civilized resistance (0.07) republic of Supreme Leader (0.07) arbitrary (0.07) separation of religion and politics (0.06) the wolves' war (0.06) blogging rallies (0.05) homosexual (0.05) green declaration (0.05) section 2 A (0.05) inviting people to demonstration (0.04) secular (0.04) mantis (0.04) Haft Tapeh Sugarcane (0.04) monopolist (0.03) body of the system (0.03) the 27th article of the Iranian constition (0.03) freedom fighters (0.03) agnostic (0.03) nationwide demonstration (0.02) addicted Ali (0.02) interrogator Hossein (0.02) the lie empire (0.02) democracy (0.02) the people in power (0.01) rotten regime (0.01) mass executions (0.01) Mansour Osanloo OR Osanloo (0.01) official governmental religion (0.01) political file (0.01) sinister regime (0.01) The reactionary velayat-faqih institution (0.01) Baha'i Faith (0.01) Varamin Prison (0) religion review (0) addicted Hossein (0) Green Silence (0) uncouth regime (0) nonviolent struggle (0) news of Bahman 25th (0) Iranian Republic (0) nonviolence (movements) (0) green media (0) civil disobedience (0)
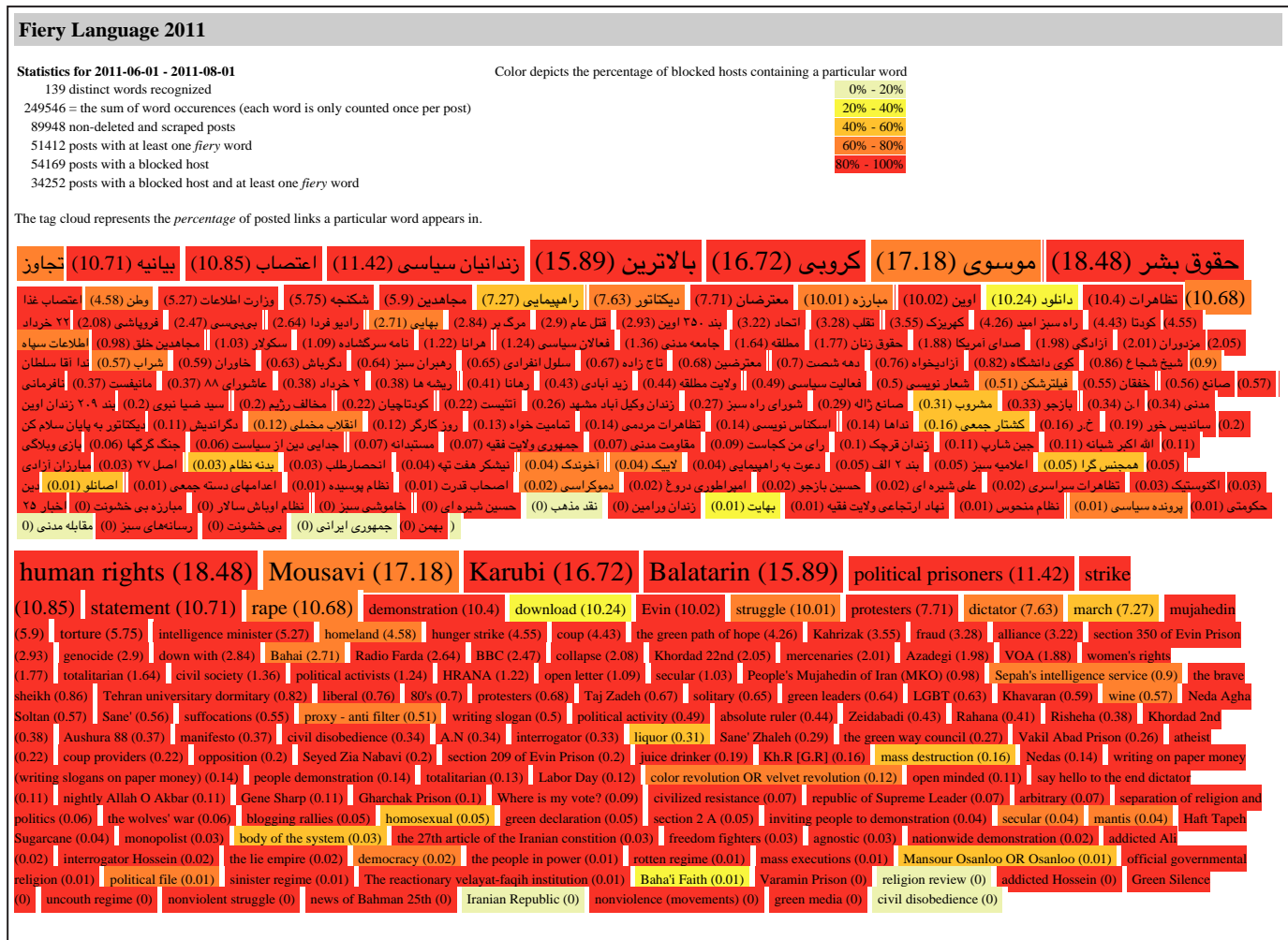
Figure 6: The "redacted web" in Iran. The use of Persian fiery language on web pages linked from Balatarin.com, June - July 2011, with English translation. The darker the color, the higher the percentage of blocked hosts containing the word. Graphic by the Digital Methods Initiative, Amsterdam. Data and additional graphics available at http://mappingiranonline.digitalmethods.net/.

2009 to 444,592 in June and July 2010 to 620,883 in June and July 2011. The use of coded language rose from 69,911 in June and July 2009 to 73,589 in June and July 2010 to 103,013 in June and July 2011.

There is a series of observations. At least for the web (and the voice) that Balatarin organizes, there is hardly a general indication of self-censorship. On the contrary, the words are to be found by censors (judging by the percentage of the same sites that have received their attention), and are in full view. We also appear to have found further indication of a hardy audience for the language, and perhaps routine censorship circumvention, if we assume that much of the readership for the fiery, side-taking and coded words is also in Iran. The use of critical language has increased, and the sites where the terms appear these days are widely blocked, showing a high rate of censorship activity and perhaps a concentration of monitoring of Balatarin. We are not able to report on specific trends in censoring sites which contain such language, as our censorship data are from August, 2011 only. Nevertheless, the overall findings are rather clear. It is a responsive web, blocked yet blogging, likely with an active readership, not only outside but also inside Iran. It would be worthwhile to collect the URLs as they pass through Balatarin (as well as Likekhor), and check for filtering simultaneously. If sites are already blocked when recommended, we have another strong indication of a culture of Internet censorship circumvention, in that there is an expectation that one is able to route around the blockage and access the sites.

Conclusion: National web health index

In this study we have sought to build upon national web characterization studies, and put forward the emerging field of national web studies. We have done so first and foremost by making a methodological plea for capturing and analyzing the diversity of national web spaces, or webs. Rather than predefining national websites, and thereby national webs, according to a principled approach of formal properties (for instance, all websites with ccTLD .ir, all websites in Persian with Iran-related content, or websites with authors inside Iran), we have concluded that such approaches are often not to be operationalized or automated. Instead we propose to make use of what we term "device cultures," and in particular the Iranian web spaces they provide, as the blogger's web, the advertiser's, the searcher's, the crowd's and the surfer's. Device cultures more specifi-

cally are defined as the interaction between user and engine, the data that are routinely collected, how they are analyzed, and ultimately the URL recommendations that result. We have demarcated national webs through devices that "go local;" they have location or language added as a value that sifts URLs that are of relevance to Iranians and Persian-speakers. In an examination of the data sets, where we performed top-level domain analysis of the sample, we have found that the majority of the collected hosts from the various Iranian webs are .com websites, not .ir, a finding that expanded the scope of national domain characterization studies, and introduced a method of data collection for a broader national web studies.

Second, both building on as well as contributing to national web characterization studies, we have proposed a rationale: a national web health index. It is conceptualized as a series of metrics, a limited number of which we have employed in this study, most readily responsiveness, page age and filtering or blockage. (We also performed language detection and top-level domain analysis.) The contribution of this work to national web characterization studies is two-fold. The first is conceptual, in that we propose to repurpose metrics from national web characterization for national web health indices. Are websites responding? Are pages fresh? Are links broken? Is the code valid? Are file formats proprietary? A form of country profiling comes into view. The second is generalizable for countries that face state censorship, and applicable to our case study in question, Iran. We compare the results from the responsiveness tests to the filtering ones. Are the blocked sites still responsive? The approach led us to find a significant number of blogs which were blocked, yet still responsive. The finding of so many blocked yet blogging sites also indicates an audience for the content, both outside Iran but also inside. We believe there to be widespread Internet circumvention in a particular space: the predominantly Persian-language blogosphere authored by Likekhor and Google Reader, which in tandem serve as an important filter for Iranian blogs. Although heavily censored, the Iranian blogosphere as listed by Likekhor remains vibrant. This censored but active space is similar to the crowd-sourced web organized by Balatarin. Blocked yet posting, Balatarin's recommended websites also suggest a similar finding as the one for the blogosphere: an active audience for blocked websites. In additional, substantive analysis we found that the Balatarin web (as a collection of URLs highly rated and thus rising

to the top of the platform) remains clamorous, perhaps even more so now than after the presidential elections of June 2009, and the initial rising of the green movement. Whilst roundly blocked, the websites comprising that Iranian web are employing critical language that is fiery, side-taking as well as coded (at least according to our three language category types we summoned for the analysis). It is a web that does not appear to be widely practicing self-censorship or one which has been cowed and drained of spirit.

Third, we would like to mention certain implications of national web studies as country profiling, both as it affects current and future policies with respect to the web (and its study) and the use of web indicators for social study more generally. As we alluded to, regarding our early work on Iraq and the state of its web during the Iraq War in 2007, national web health study provides an additional set of measures regarding the current state of the universities, ministries and other institutions. Where is the activity, and where is the neglect? It also may serve as a source of comparative study, and ultimately as a spur to addressing the ill health of one or more webs. Thus it is an approach to the study of the web that could have salutary consequences for portions of it.

References

Abiteboul, S., Cobena, G., Masanès, J., and Sedrati, G. (2002). "A first experience in archiving the French Web." Paper presented at the 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, 16-18 September.

Amir-Ebrahimi, M. (2008). "Blogging from Qom, behind Walls and Veils," *Comparative Studies of South Asia, Africa and the Middle East*. 28(2): 235-249.

Anderson, B. (1991). *Imagined Communities*. London: Verso.

Arms, W. Y., Adkins, R., Ammen, C., and Hayes, A. (2001). "Collecting and preserving the Web: The Minerva prototype." *RLG DigiNews*. 5(2). <http://www .rlg.org/preser v/diginews/diginews5-2.html> [accessed 9 September 2011].

Arvidson, A. and Lettenström, F. (1998) "The Kulturarw Project - The Swedish Royal Web Archive, *The Electronic Library*. 16(2): 105-108.

Baeza-Yates, R., Castillo, C. and Efthimiadis, E.N. (2007). "Characterization of National Web Domains," *Journal ACM Transactions on Internet Technology*. 7(2), Art. 9.

Blogpulse (2011). "FAQ: How do you determine blog rankings?", Blogpulse.com <http://blogpulse.com/about.html#profiles_3> [accessed 9 September 2011]

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage*. New York: Peter Lang.

Deibert, R. and Rohozinski, R. (2010). "Cyber wars," *Index on Censorship*. 29(1): 79-90.

Deuze, M. (2007). *Media Work*. Cambridge: Polity.

Drummond, D. (2010a) "A New Approach to China." *The Official Google Blog*. 12 January. <http://googleblog.blogspot.com/2010/01/new-approach-to-china.html> [accessed 9 September 2011].

Drummond, D. (2010b) "An update on China." *The Official GoogleBlog*. June 28. <http://googleblog.blogspot.com/2010/06/update-on-china.html> [accessed 9 September 2011].

Etling, B., Alexanyan, K., Kelly, J., Faris, R., Palfrey, J. and Gasser, U. (2010). "Public Discourse in the Russian Blogosphere: Mapping RuNet Politics and Mobilization," Berkman Center Research Publication No. 2010-11. <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Public_Discourse_in_the_Russian_Blogosphere_2010.pdf> [accessed 9 September 2011].

Feuz, M., Fuller, M. and Stalder, F. (2011). "Personal Web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalization," *First Monday*. 16(2). <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3344/2766> [accessed 9 September 2011].

Gerlitz, C. and Helmond, A. (2011) "The Like economy: the social web in transition," Paper presented at the MIT7 Unstable Platforms conference, MIT, Cambridge, MA, 13-15 May.

Ginsburg, F., Abu-Lughod, L. and Larkin, B. (2002). "Introduction." in: Ginsburg, F., Abu-Lughod, L. and Larkin, B. (eds.). *Media Worlds: Anthropology on new terrain*. Berkeley: University of California Press.

Glanz, J. and Markoff, J. (2011) "U.S. Underwrites Internet Detour Around Censors," *New York Times*, 12 June. <https://www.nytimes.com/2011/06/12/world/12internet.html> [accessed 9 September 2011].

Goldsmith, J. and Wu, T. (2006) *Who Controls the Internet. Illusions of a Borderless World*. Oxford: Oxford University Press.

Google (2011a). "Company," website. <http://www.google.com/about/corporate/company/> [accessed 9 September 2011].

Google (2011b). "Google Transparency Report," website.

<http://www.google.com/transparencyreport/> [accessed 25 August 2011].

Higson, A. (1989). "The Concept of National Cinema," *Screen*. 30(4): 36-47.

Howe, J. (2006). "The Rise of Crowdsourcing," *Wired*. 14(6). <http://www.wired.com/wired/archive/14.06/crowds.html> [accessed 9 September 2011].

International Telecommunication Union (2011), *Measuring the Information Society*. ITU: Geneva.

Iran Media Program (2010). "Balatarin: a battleground for defining freedom of expression," Blog. Annenberg School for Communication, University of Pennsylvania. <http://iranmediaresearch.org/en/blog/13/10/11/23/201> [accessed 9 September 2011].

Kehoe, C., Pitkow, J., Sutton, K., Aggarwal, G. & Rogers, J. (1999). "GVU's Tenth World Wide Web User Survey," Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology, Atlanta, Georgia.

Kelly, J. and Etling, B. (2008). "Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere," Berkman Center Research Publication No. 2008-01. <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Kelly&Etling_Mapping_Irans_Online_Public_2008.pdf> [accessed 9 September 2011].

Khiabany, G. and Sreberny, A. (2007). "The Politics of/in Blogging in Iran," *Comparative Studies of South Asia, Africa and the Middle East*. 27(3): 563-579.

Koerbin, P. (2004). "The Pandora Digital Archiving System (PANDAS) and managing Web archiving in Australia: A case study." Paper presented at the 4th International Web Archiving Workshop, Bath (UK), September 16.

Lasfargues, F., Oury, C., and Wendland, B. (2008) "Legal Deposit of the French Web: Harvesting Strategies for a National Domain," Paper presented at IWAW'08, 18-19 September, Aarhus, Denmark. <http://iwaw.europarchive.org/08/IWAW2008-Lasfargues.pdf> [accessed 9 September 2011].

Miller, D. and Slater, D. (2000). *The Internet: An Ethnographic Approach*. Oxford: Berg.

Moulier-Boutang, Y. (2008). "Worker Bee Economy." Paper presented at the Society of the Query Conference, Institute of Network Cultures, Amsterdam.

Murdoch, S. and Anderson, R. (2008). "Tools and Technology of Internet Filtering," In Deibert, R., Palfrey, J., Rohozinski, R. and Zittrain, J. (eds.), *Access Denied: The Practice and Policy of Global Internet Filtering*. Cambridge, MA: MIT Press.

MVF Global (2010), "Online Marketing in the top 50 Internet Economies: Lead Generation and Internet Marketing in Iran." <http://www.mvfglobal.com/iran> [accessed 20 October 2011].

Net Applications (2011), "Search Engine Market Share: Iran, Islamic Republic of." <https://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4&qpaf= 000%09101%09IR%0D&qptimeframe=Y> [accessed 20 October 2011].

NetBina (2010), "Online Marketing in Iran." http://new.netbina.com/resources/2/docs/Online_marketing_in_Iran_2010.pdf

Open Net Initiative (2005). "Internet Filtering in Iran in 2004-2005: A Country Study," Toronto: University of Toronto. <http://opennet.net/studies/iran> [accessed 9 September 2011].

Open Net Initiative (2009). "Internet Filtering in Iran, 2009," Toronto: University of Toronto. http://opennet.net/sites/opennet.net/files/ONI_Iran_2009.pdf [accessed 9 September 2011].

PADI (n.d.). "Legal Deposit," Preserving Access to Digital Information initiative, National Library of Australia. <https://www.nla.gov.au/padi/topics/67.html> [accessed 9 September 2011].

Pariser, E. (2011). *The Filter Bubble*. New York: Penguin.

Quenqua, D. (2009). "Blogs Falling in an Empty Forest," *New York Times*, 5 June. <http://www.nytimes.com/2009/06/07/fashion/07blogs.html> [accessed 9 September 2011].

Rhoads, C. and Fassihi, F. (2011). "Iran Vows to Unplug Internet," *Wall Street Journal*, 28 May. <http://online.wsj.com/article/SB10001424052748704889404576277391449002016.html> [accessed 20 October 2011].

Roberts, H., Zuckerman, E. and Palfrey J. (2011). "2011 Circumvention Tool Evaluation," Berkman Center for Internet and Society, August. <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/2011_Circumvention_Tool_Evaluation_1.pdf> [accessed 9 September 2011].

Rogers, R. (2002). "Operating Issue Networks on the Web," *Science as Culture.* 11(2): 191-214.

Rogers, R. (2009). *The End of the Virtual: Digital Methods*. Amsterdam: Amsterdam University Press.

Schmidt, E. (2009). "Prosperity or Peril? The Next Phase of Globalization," Princeton Colloquium on Public and International Affairs, 18 April. <http://www.youtube.com/watch?v=9nXmDxf7D_g> [accessed 9 September 2011].

Technorati (2011). "Blog Quality Guidelines," Technorati.com, <http://technorati.com/blog-quality-guidelines-faq> [accessed 9 September 2011]

Thompson, C. (2006). "Google's China Problem (and China's Google Problem)," *New York Times*, 23 April. <http://www.nytimes.com/2006/04/23/magazine/23google.html> [accessed 9 September 2011].

Villeneuve, N. (2006). "Testing through proxies in China," Nart Villeneuve blog, 10 April. <http://www.nartv.org/2006/04/10/testing-through-proxies-in-china/> [accessed 9 September 2011].

Weltevrede, E. (2009). "Thinking Nationally with the Web: A Medium-Specific Approach to the National Turn in Web Archiving." M.A thesis, University of Amsterdam.

Whetstone, R. (2010) "Controversial content and free expression on the web: a refresher." *The Official Google Blog.* April 19. <http://googleblog.blogspot.com/2010/04/controversial-content-and-free.html> [accessed 9 September 2011].

Wikipedia (2011). "Balatarin." 23 July. <http://en.wikipedia.org/wiki/Balatarin> [accessed 9 September 2011].

Wright, J., de Souza, T. and Brown, I. (2011) Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics. FOCI'11 (USENIX Security Symposium), 8 August 2011, San Francisco. <http://www.usenix.org/events/foci11/tech/final_files/Wright.pdf> [accessed 9 September 2011].

Yossef, Z. B., Broder, A. Z., Kumar, R., and Tomkins, A. (2004). "Sic transit gloria telae: towards an understanding of the web's decay." *Proceedings of the 13th conference on World Wide Web*. New York: ACM.

Zarrinbakhsh, N. (2011). "Living as a criminal: An ethnographic study of the Iranian national web and Internet censorship." M.A. thesis, University of Amsterdam.

Appendix

Persian queries	English translation	Number of posts in which the Persian word appears	fiery	side taking	code language
خرداد ۲	Khordad 2nd	652	1	0	0
خرداد ۲۲	Khordad 22nd	4723	1	0	0
بهمن ۲۵	Bahman 25th	2132	0	1	0
آتئیست	atheist	291	1	1	0
آخوندک	mantis	180	1	0	1
آزادگی	Azadegi	3413	1	0	0
آزادی بیان	freedom of speech	3837	0	1	0
آزادیخواه	liberal	1811	1	1	0
آهنگ	music	13415	0	1	0
ا.ن	A.N.	2198	1	0	1
اتحاد	alliance	9664	1	1	0
اتوبوس رانی	bus drivers union	252	0	1	0
اخبار ۲۵ بهمن	news of Bahman 25th	1	1	1	0
استراتژی	strategy	3911	0	1	0
اسکناس نویسی	writing on paper money (writing slogans on paper money)	370	1	0	1
اصانلو	Mansour Osanloo OR Osanloo	22	1	0	0
اصحاب قدرت	the people in power	65	1	1	1
اصل ۲۷	the 27th article of the Iranian	198	1	0	1
اصل ولایت فقیه	the principle of Guardianship of the Islamic Jurists (velayat-e faqih)	714	0	1	0
اصلاحات	reforms	10960	0	1	0
اطلاع رسانی	spreading information	12447	0	1	0
اطلاعات سپاه	Sepah's intelligence service	1326	1	0	0
اعتصاب	strike	17034	1	1	0
اعتصاب غذا	hunger strike	6503	1	1	0
اعدام	execution	20597	0	1	0
اعدامهای دسته جمعی	mass executions	17	1	1	0
اعلامیه	declaration	3795	0	1	0
اعلامیه سبز	green declaration	61	1	0	0
اقتصادی	economic	35384	0	1	0
اگنوستیک	agnostic	77	1	0	0
الله اکبر شبانه	nightly Allah O Akbar	238	1	0	1
امپراطوری دروغ	empire of lies	32	1	0	0
انتخابات	election	49469	0	1	0
انحصارطلب	monopolist	79	1	1	0
انقلاب	revolution	51257	0	1	0
انقلاب مخملی	color revolution OR velvet revolution	834	1	0	0
اوین	Evin	15630	1	1	0
بازجو	interrogator	903	1	1	0
بازداشت	arrest	33211	0	1	0
بازی وبلاگی	blogging rallies	366	1	0	1
بالاترین	Balatarin	28890	1	1	1
بحران	crises	18830	0	1	0
بدنه نظام	body of the system	54	1	1	1
بند ۲ الف	section 2 A	98	1	0	0
بند ۲۰۹ زندان اوین	section 209 of Evin Prison	343	1	0	0
بند ۳۵۰ اوین	section 350 of Evin Prison	2987	1	0	0
به پا خاستن	uprising	38	0	1	0

بهایت	Baha'i Faith	8	1	0	0
بهایی	Baha'i	3698	1	0	0
بهمن سال ۸۹	January 2011	13	0	1	0
بی خشونت	nonviolence (movements)	8	1	1	0
بیانیه	statement	20929	1	1	0
بی‌بی‌سی	BBC	5361	1	0	0
پرونده سیاسی	political file	18	1	1	0
پوستر	poster	3557	0	1	0
تاج زاده	Taj Zadeh	2754	1	0	0
تاکتیک	tactic	1938	0	1	0
تجاوز	rape	17211	1	0	0
تجمعات	gatherings	3671	0	1	0
تشکیلات	organizations	2321	0	1	0
تصرف	occupation	6618	0	1	0
تصرف صدا و سیما	occupation of the national TV (IRIB)	3	1	0	0
تظاهرات	demonstration	23143	1	1	0
تظاهرات سراسری	nationwide demonstration	79	1	1	0
تظاهرات مردمی	people demonstration	351	1	1	0
تقلب	fraud	10336	1	1	0
تقلب در انتخاب	elections fraud	5	1	0	0
تمامیت خواه	totalitarian	312	1	1	0
تمساح	alligator	903	0	0	1
تونس	Tunisia	3851	0	1	0
جامعه	society	49036	0	1	0
جامعه مدنی	civil society	1945	1	1	0
جدایی دین از سیاست	separation of religion and politics	242	1	1	0
جزوه	leaflet	800	0	1	0
جمهوری ایرانی	Iranian Republic	15	1	0	1
جمهوری خواهی	republicanism	4	1	1	0
جمهوری ولایت فقیه	republic of Supreme Leader	99	1	0	1
جنبش	movement	35113	0	1	0
جنگ گرگها	the wolves' war	65	1	1	1
جین شارپ	Gene Sharp	151	1	0	0
حاکمیت	rule	9752	0	1	0
حبس خانگی	house arrest	2590	0	1	0
حسین بازجو	interrogator Hossein	83	1	0	1
حسین شوش	Hossein Shoosh	4	1	0	1
حسین شیره ای	addicted Hossein	5	1	0	1
حقوق بشر	human rights	33448	1	1	0
حقوق زنان	women's rights	3262	1	1	0
حکم	sentence	24191	0	1	0
خ.ر	Kh.R [G.R]	752	1	0	1
خاتمی	Khatami	26151	0	1	0
خاموشی سبز	Green Silence	209	1	0	0
خاوران	Khavaran	961	1	1	1
خرداد	Khordad	54343	0	1	0
خرید عید	Nowruz shopping	50	0	0	1
خس و خاشاک	a pile of dust	3465	0	0	1
خشم	anger	6782	0	1	0
خفقان	suffocations	2862	1	1	0
دانلود	download	19786	1	1	0
درگیری	fighting	15677	0	1	0
دستگیری	arrest	15318	0	1	0

دعوت	invite	21793	0	1	0
دعوت به راهپیمایی	inviting people to demonstration	77	1	0	0
دفاع	defend	31866	0	1	0
دگراندیش	open minded	250	1	0	0
دگرباش	LGBT	718	1	0	0
دموکراسی	democracy	20	1	1	0
دهه شصت	80's	1772	1	0	0
دیکتاتور	dictator	12677	1	0	1
دیکتاتور به پایان سلام کن	the end is near, dictator	112	1	0	1
دین حکومتی	official governmental religion	19	1	0	0
رادیو فردا	Radio Farda	6394	1	0	0
راه سبز امید	the green path of hope	6749	1	0	0
راهپیمایی	march	13952	1	0	0
رای من کجاست	Where is my vote?	421	1	0	0
رسانه‌های سبز	green media	1	1	0	0
رهانا	Rahana	1035	1	0	0
رهبران سبز	green leaders	1069	1	0	1
روز جهانی زن	intl. Women's Day	1	1	0	0
روز کارگر	Labor Day	390	1	0	0
ریشه ها	Risheha	696	1	0	1
زباله دانی تاریخ	dustbin of history	59	0	0	1
زندان	prison	37708	0	1	0
زندان قرچک	Gharchak Prison	125	1	0	0
زندان ورامین	Varamin Prison	5	1	0	0
زندان وکیل آباد مشهد	Vakil Abad Prison	413	1	0	0
زندانی	prisoner	23505	0	1	0
زندانیان سیاسی	political prisoners	17225	1	0	0
زید آبادی	Zeidabadi	704	1	0	0
ساندیس خور	juice drinker	395	1	0	1
سبز	green	43200	0	1	1
سرپیچی	refusal	494	0	1	0
سرکوب	suppression	18260	0	1	0
سطل زباله	trash bin	281	0	1	1
سقوط	collapse	25283	0	1	0
سکولار	secular	2483	1	0	0
سلول انفرادی	solitary	1384	1	0	0
سه شنبه‌های اعتراض	the Tuesday protests	64	1	0	0
سهراب	Sohrab	8207	0	1	1
سیاسی	political	57438	0	1	0
سید ضیا نبوی	Seyed Zia Nabavi	361	1	1	0
شبکه اجتماعی	social network	2149	0	1	0
شراب	wine	1657	1	0	0
شرایط بحرانی	crisis conditions	540	0	1	0
شرایط ناگوار	terrible conditions	31	0	1	0
شعار	slogan	18340	0	1	0
شعار نویسی	writing slogan	972	1	0	0
شکنجه	torture	12450	1	0	0
شهادت	martyrdom	15885	0	1	0
شورا	council	3869	0	1	0
شورای راه سبز	the green way council	269	1	0	0
شیخ شجاع	the brave sheikh	929	1	0	1
صانع	Sane'	983	1	0	1
صانع ژاله	Sane' Zhaleh	653	1	0	0

صدای آمریکا	VOA	4379	1	0	0
صندوق رای	voting box	510	0	1	0
ضرب و شتم	beating	8278	0	1	0
ظلم و جور	oppression	367	0	1	0
عاشورای ۸۸	Aushura 88	459	1	0	0
عدالت اجتماعی	social justice	1369	0	1	0
علوم انسانی	humanities	5	0	1	0
علی شیره ای	addicted Ali	105	1	0	1
غیر مسلحانه	unarmed	15	0	1	0
فروپاشی	collapse	3204	1	1	0
فشار اجتماعی	social pressure	33	0	1	0
فعالان سیاسی	political activists	3138	1	0	0
فعالیت سیاسی	political activity	1191	1	1	0
فعالین	activists	5912	0	1	0
فیلترشکن	proxy-server as anti-filtering	1193	1	0	0
فیلم	video clip	38021	0	1	0
قتل عام	genocide	5156	1	0	0
قرنطینه	quarantine	487	0	1	0
قزل قلعه	Ghezal Ghaleh	40	0	1	1
کذاب	liar	298	0	1	1
کروبی	Karrubi	30413	1	0	0
کشتار	massacre	10167	0	1	1
کشتار جمعی	mass destruction	344	1	1	0
کمپین	campaign	8687	0	1	0
کهریزک	Kahrizak	6537	1	0	1
کودتا	coup	10461	1	0	1
کودتاچیان	coup providers	1791	1	0	1
کوی دانشگاه	Tehran universitary dormitary	3233	1	0	0
گزارش	report	72906	0	1	0
گفتمان	discourse	3430	0	1	1
لاییک	secular	86	1	0	0
لیبرال	liberal	1950	0	1	0
مانیفست	manifesto	855	1	0	0
مبارزان آزادی	freedom fighters	36	1	0	1
مبارزه	struggle	23879	1	1	0
مبارزه بی خشونت	nonviolent struggle	1	1	0	1
مجاهدین خلق	People's Mujahedin of Iran (MKO)	1167	1	0	0
مجاهدین	mujahedin	9895	1	1	0
محکومیت	sentence	7629	0	1	0
مخالف رژیم	opposition	335	1	0	0
مختاری	Mokhtari	1276	0	1	0
مرگ بر	down with	8837	1	0	0
مزدوران	mercenaries	3883	1	0	1
مستبدانه	arbitrary	187	1	0	0
مسلحانه	armed	3041	0	1	0
مشروب	liquor	651	1	0	0
مشروعیت	legitimacy	4579	0	1	0
مصر	Egypt	14551	0	1	1
مطالبات مردم	people's demands	674	0	1	0
مطبوعاتی	press	5253	0	1	0
مطلقه	totalitarian	4135	1	0	1
معترضان	protesters	13326	1	0	0
معترضین	protesters	2285	1	0	0

مقابله مدنی	civil disobedience	4	1	0	0
مقاومت	resistance	14843	0	1	0
مقاومت مدنی	civilized resistance	237	1	1	0
مناظره	debate	6663	0	1	1
موسوی	Mousavi	42838	1	1	0
میهن	homeland	6563	0	1	0
نارضایتی	discontent	3776	0	1	0
نافرمانی	disobedience	2735	0	1	0
نافرمانی مدنی	civil disobedience	1903	1	1	0
نامه	letter	54699	0	1	1
نامه سرگشاده	open letter	3734	1	1	0
ندا	Neda	11014	0	1	1
ندا آقا سلطان	Neda Agha Soltan	2458	1	1	0
نداها	Nedas	474	1	0	1
نظام اوباش سالار	uncouth regime	2	1	0	1
نظام پوسیده	rotten regime	29	1	0	1
نظام منحوس	sinister regime	14	1	0	1
نقد مذهب	religion review	17	1	1	0
نهاد ارتجاعی ولایت فقیه	The reactionary velayat-faqih	6	1	0	1
نیشکر هفت تپه	Haft Tapeh Sugar Cane	74	1	1	0
هرانا	HRANA	2280	1	0	0
همبستگی	solidarity	119	0	1	0
همجنس گرا	homosexual	171	1	0	0
وزارت اطلاعات	intelligence minister	10012	1	1	0
وطن	homeland	10173	1	1	0
ولایت مطلقه	absolute ruler	1046	1	0	0
یا حسین	Ya Hossein (2nd Shia Imam name)	1334	0	1	1

Screenshot of block page, http://10.10.34.34, indicating that the web user in Iran has attempted to access a banned website. Block page by the Ministry of Culture and Islamic Guidance, Iran, February, 2012, captured by the Censorship Explorer tool by the Digital Methods Initiative, Amsterdam.