Differentiations in Internet Access: Problems and Challenges[1]

Tyng-Ruey Chuang
Institute of Information Science
Academia Sinica

It was about 10 years ago Tim Berners-Lee, the inventor of World Wide Web, first programmed a computer in CERN, a European high-energy research organization, to serve Web pages over the Internet. Internet's origin, however, can be traced far back to the 1960's, when the US defense research project office DARPA started funding research on interconnecting computer networks. The Web did not take off until 1993, the year "dotcom" web sites started to appear, and Mosaic, one of the first Web browsers, was released. The Internet has changed a lot since then. The number of Internet users has increased 1,000-fold, from just a few hundred thousands, mostly in Western universities, to hundreds of millions, worldwide. Along with the growth of Internet population, Web-based systems have been proliferating to move online various social and business activities. Because the "Internet time" has been moving so fast, and the ensuing change so rapid, we may have yet to fully understand the Web's full impact to the society. In this paper, we discuss factors that differentiate people's access to the Web, based on observations drawn from readily available population data and Web statistics.

Does the popularity of Web help reduce people's differences, at least in their ability to access information? Or is there an enlarging gap, so-called the "digital divide," that separates the digital have from the digital have-not? These basic questions still elude definite answers. By

[1] Presented at the *Pacific Neighborhood Consortium (PNC) Annual Conference and Joint Meetings*, 15-20 January 2001, Hong Kong, as part of the Internet Access and Censorship session. We thank the conference organizers and the session chair for the invitation. I thank Mr. Jan-Li Lin, my research assistant at the Institute of Information Science, Academia Sinica, for his help in collecting the data and preparing the illustrations.

looking into the demographic distributions of Internet users, host machines, and Web pages — as we will do shortly — one can locate populations that are information rich as well as those that are information poor. In this paper, we hope to highlight the disparities between the two groups in their access to the Internet. We also outline the capacity of Internet to provide differentiated Web services, and describe how web users can be, and often are, served differently even if they are connected to the same Web site. These "differentiations in Internet access" may not catch the attention of many Internet users, but are the obstacles to the realization of a truly online society.

By one estimation, there are about 380 million Internet users today, roughly one-sixteenth of the entire world population[2]. About 190 million English speakers are online, that is 60% of the entire English-speaking population (320 millions), and one-half of the entire online population. On the contrary, of the 1.2 billion Chinese-speaking population, less than 30 millions are Internet users. The Spanish-speaking population is about ten millions more than the English-speaking population, but its online population is only 19 millions, one-tenth of the English online population.[3] So there is already significant disparity in the distribution of Internet users if one takes into account the languages they use. When one classifies Web pages on the Internet by the languages they are written with, one finds greater disparity. Of the 1.6 billion Web pages crawled by Google, a search engine company, 1.2 billions are in English. That is 75% of the entire collection. However, only 27 million pages are in Chinese and 13 million pages in Spanish: They translate into puny 1.7% and 0.8% of the total.[4] Clearly, when compared to people who do not speak English, English speakers are more likely to have Internet access and, once on the net, have access to far more Web pages. More details can be found in Table 1, where we include the related numbers for the 15 most popular languages, as well as some languages that have strong Web presence. We also observe the lack of data about online

---

[2] Estimated by Global Reach (http://www.glreach.com) as of September 2000.

[3] Populations of Internet users, by the primary languages they speak, are from Global Reach as estimated in September 200. World populations, also by primary languages, are from SIL International (http://www.sil.org) as estimated in February 1999.

[4] These numbers are from Google, Inc. (http://www.google.com), as cited in the *Wired* magazine, December 2000.

population and Web pages for several of the most popular languages used in South Asia, like Bengali, Hindi, Javanese, Vietnamese, and Telugu. The lack itself is an indication of these languages' weak Web presence and their insignificance in the global Internet society.

One can also characterize the differentiations in access by Internet domain names, which are associated to host computers on the Internet. According to Google, 60% of the 1.6 billion Web pages come from three top-level domains: .com, .net, and .org. Web pages from these "generic top-level domains" (gTLD) can be served by machines in any country (though many are located in US), and can be thought to provide information service to the world-wide population (though the Web pages may mostly be in English). Another 33% of the Web pages are from the "country code top-level domains" (ccTLD), like .hk (Hong Kong). Though they are equally accessible to all Internet users, Web pages from a specific ccTLD are primarily targeted to Internet users in the corresponding country. Table 2 lists the population, the number of host machines on the Internet, and the number of accessible Web pages for each of the 15 most populated ccTLDs, as well as some ccTLDs that have strong Web presence. One cannot help but notice significant differences among the countries. For example, the population of Japan is about one-tenth of China, but the number of Internet hosts in Japan is 40 times over that in China, and the number of Web pages in Japan is 7 times over that of China. Notice that English is not the primary languages in both countries. Hence, one can view the numbers of host machines and Web pages as indicators of how well the Internet is used by the people in the country to communicate with one another. If one divides these numbers by the population of the country, one find out that, per person, the density of Internet hosts in Japan is 400 times over those in China, and the density of Web pages in Japan is 70 times over that in China. Similarly, if one compares Taiwan to Japan, the Internet host density in Taiwan is 1.5 times over that in Japan, and the Web page density in Taiwan is 1.6 over that in Japan. Again, if one compares Demark to Taiwan, the ratios are 1.7 for host density and 2.3 for page density. Table 2 also reveals that Nigeria (population 110 millions) and Vietnam (population 82 millions) each has only about 100 Internet hosts. From

Table 1 and Table 2, we can only conclude that the languages one uses, and the country one resides, often decide whether one is online, and, if online, the quantity of Web pages one has access to. We further illustrate the differences using pie charts in Figure 1 and Figure 2.

For people who speak the same language and live in the same country, the distribution of online population can be as deviated as if they were from different countries. Income and education levels usually are the deciding factors. If one accesses the Internet at home, then one's household income must be able to support the purchase of a home PC and to pay the fee for Internet access fee. If one accesses the Internet at work, then one probably has some PC-related work skill, and is equipped with the necessary training or education. Even for people who are online, they may experience the Web in quite different ways. For example, for people who have broadband access, the bandwidth will allow them to browse video-rich content more often. For people who use cell phones to access the Internet, then most likely they are looking up specific information rather than browsing the Web for fun. In particular, for people with disability, *e.g.*, the blind, they can only access Internet with the assistance of special software, and can only browse Web pages that are properly prepared.

The commercialization of Internet also poses additional problems, in that content providers, for business considerations, are increasingly interested in categorizing users who visit their Web sites. By design, the Web does provide the necessary mechanism for them to perform the categorization. It may not be evident to many users, but the simple act of browsing a Web site actually is the outcome of a sequence of negotiations between two pieces of software, the user's Web browser and the content provider's Web server. The Web browser usually reveals a lot about its user. For example, it sends over the user's preferred language setting, and the user machine's "IP address", in order to receive information from the Web server. The IP address uniquely identifies one machine within the entire Internet, and can be related to the machine's geographic area. (One can think of it as one's telephone number, completed with the country and area code.) There are also "cookies", which are pieces of information sent by the

browser to the Web server, or to a group of Web servers, about whether or not a user has been using their sites, and the kinds of content the user prefers. In summary, users' access to the Internet may not be anonymous. This is rather different from access to traditional media like newspaper or broadcasting, which is almost always anonymous.

This loss of anonymity also reflects in the third party's ability to disrupt or condition the flow of information between the user's Web browser and the content provider's Web server. A third party, because of its control of critical routes in the network, can regulate information flow on the routes based on the IP addresses of the users and/or the content providers. Some content providers, in order to avoid this disruption, may even cooperate with the third party, often a government, in delivering customized Web pages to users in specific areas. As a result, Internet users are not served equally, even if they are connected to the same Web site.

There is also a great inequality among the Internet users on how well they can get information from the entire Web. Commercial users, like companies specialized in Web search, can best harvest the Internet because they are possessed of the necessary computational resource and technical know-how. Actually, ordinary Internet users depend on these commercial Web users for information on the Web. This may not be a problem, but again it illustrates that not all Internet users are equal: Some are more powerful than others. People also depend on popular Web sites (like AOL and Yahoo) for online forums (*e.g.*, BBS and mailing lists), and as a result a mass amount of social activities has being recorded at these sites. As a consequence, not only do the services they provide become indispensable to people's daily activities, but the collection of records the users left on these sites become a significant part of the human history. We think it will soon be a pressing issue on how these records should be archived (if they should be archived at all) for the future historians. Notice that for people who are not online today, they leave behind no record on the Web, and will not be represented in the historical archive of tomorrow.

We have shown that there are great disparities in the distribution of Internet population, and in the amount of Web pages people have access to. We also briefly illustrate why and how some Web sites provide differentiated services to their users. How do us face these disparities, and what are the remedies? We cannot help but ask these questions, knowing full well that there will be no ready answer. For poor countries, we feel that economy development and education level are the most decisive factors in taking more people online. For countries that already have sizable Internet population, many of the poor and the under-educated will not get online if their economical or educational status remains unchanged. The government needs to play a more active role in bringing Internet access to people. For example, the government can set up a "universal access" mechanism to make sure that telecommunication companies and Internet service providers do meet the needs of the poor, and do provide special assistance to people with disability.

In April 1999, Tim Berners-Lee gave a speech on "The Future of the Web". He used the term "social machines" to refer to the Web's roles in bringing people to work together and to make progress.[5] Internet and the Web indeed make incredible social machines. Without them, there will be no mailing lists, no newsgroups, no online sharing of data (*e.g.*, napster), and no open source software development, just to name a few. For people who do not have access to these social machines, they cannot participate, and will not be members of the society.

[5] Tim Berners-Lee, *The Future of the Web*, speech at the 35th anniversary of MIT Laboratory For Computer Science, April 14, 1999 (http://www.w3.org/Talks/1999/0412-LCS35-tbl).

Table 1. Distribution of population, Internet users, and Web pages (by languages).

Language	Population	Internet Users	Web Pages
Chinese	1223.3	28.7	27.0
Spanish	332.0	18.9	13.0
English	322.0	189.6	1225.4
Bengali	189.0	--	--
Hindi	182.0	--	--
Portuguese	170.0	9.4	5.6
Russian	170.0	9.3	--
Arabic	167.0	2.4	--
Japanese	125.0	27.3	44.3
German	98.0	23.4	36.5
Javanese	75.5	--	--
Korean	75.0	15.7	10.4
French	72.0	15.2	17.4
Vietnamese	68.0	--	--
Telugu	66.4	--	--
Italian	37.0	12.3	9.9
Dutch	20.0	6.0	5.8
Swedish	9.0	3.6	5.1
Finnish	6.0	2.2	2.7
Danish	5.3	2.6	2.2
Norwegian	5.0	2.2	2.2
Hebrew	4.6	1.0	0.3
Icelandic	0.3	0.1	0.3
Others	2212.5	12.1	191.7
Total	5634.9	382.0	1600.0

Unit: million; '--': not available.
Sources: SIL International (population, as of February 1999, http://www.sil.org), Global Reach (Internet users, as of September 2000, http://www.glreach.com), and Google (Web pages, as reported in *Wired*, December 2000).

Table 2. Distribution of population, Internet hosts, and Web pages (by domains).

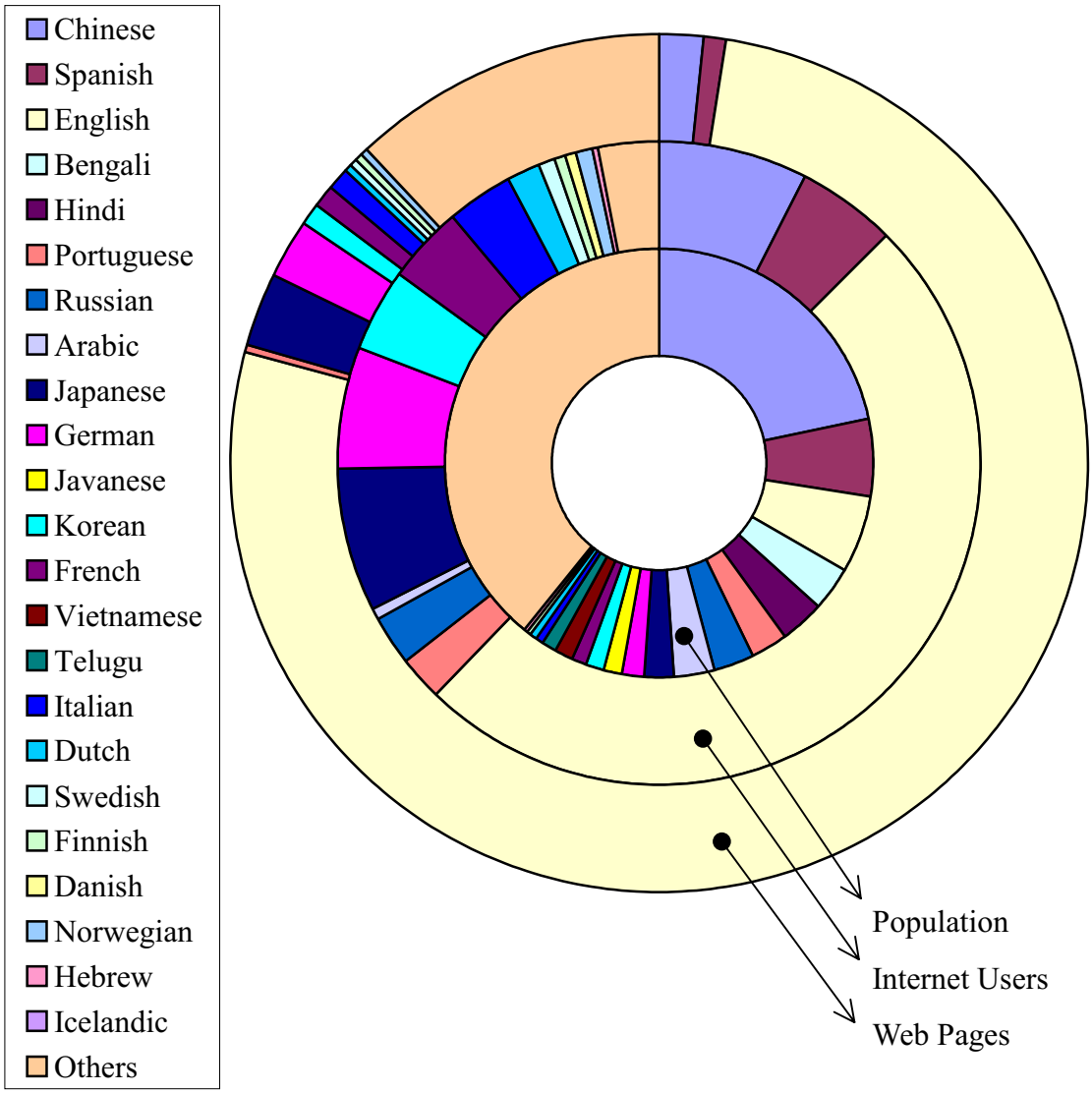
Country	Domai	Populat	Internet	Web
(Commerce)	.com	--	32696.3	783.0
(Network)	.net	--	23432.1	101.4
(Education,	.edu	--	6678.1	81.3
(Organization)	.org	--	1087.7	76.2
(Government,	.gov	--	827.6	10.1
China	.cn	1278.	87.9	10.9
India	.in	1017.	33.0	--
United States	.us	278.1	2251.4	10.2
Indonesia	.id	214.4	24.1	--
Brazil	.br	167.6	662.9	--
Russian	.ru	145.1	260.4	10.7
Pakistan	.pk	141.5	5.5	--
Japan	.jp	127.9	3413.3	69.0
Nigeria	.ng	114.3	0.1	--
Mexico	.mx	98.8	495.7	--
Germany	.de	82.9	1916.5	76.5
Vietnam	.vn	82.2	0.1	--
Philippines	.ph	80.1	16.7	--
Turkey	.tr	66.2	108.4	--
Thailand	.th	63.1	53.7	--
France	.fr	59.7	983.5	16.0
United	.uk	59.6	2080.9	64.2
Italy	.it	57.9	1574.4	15.5
South Korea	.kr	47.7	475.8	11.5
Spain	.es	39.6	538.5	11.4
Canada	.ca	31.2	1814.5	12.5
Taiwan	.tw	22.3	903.9	19.2
Australia	.au	19.4	1311.5	10.4
Netherlands	.nl	15.9	1082.1	17.6
Sweden	.se	8.9	624.3	11.8
Switzerland	.ch	7.3	418.0	13.0
Denmark	.dk	5.4	369.7	10.7
Others	--	1723.	6819.3	157.0
Total	--	6055.	93047.8	1600.0

Unit: million (population, Web pages), thousand (Internet hosts); '--' : not available.

Sources: The Economist (population, *The World in 2001*), Internet

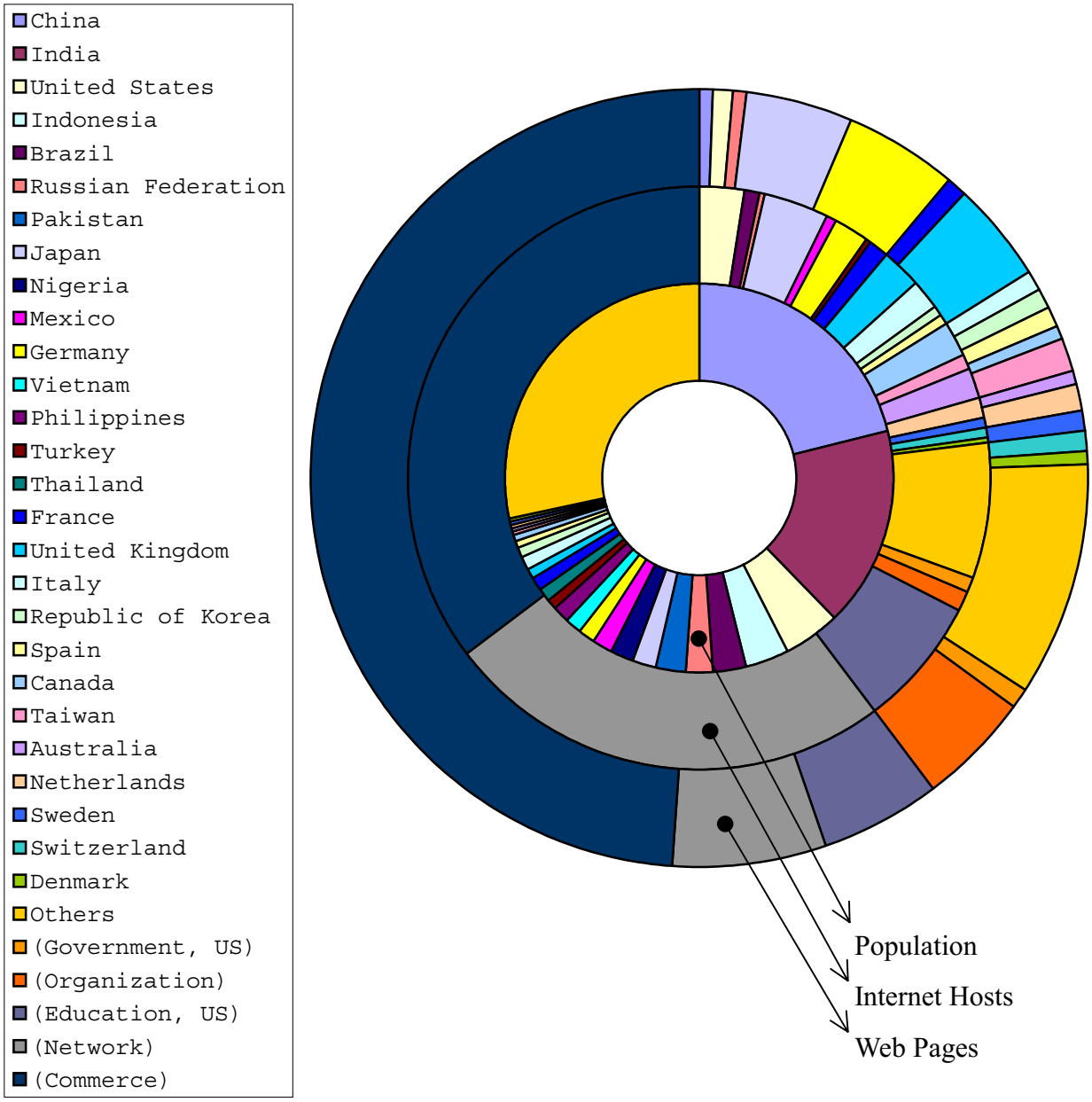
Software Consortium (Internet hosts, as of July 2000, http://www.isc.org), and Google (Web pages, as reported in *Wired*, December 2000).

Figure 1. Distribution of population, Internet users, and Web pages (by languages).



Sources: SIL International (population), Global Reach (Internet users), and Google (Web pages).

Figure 2. Distribution of population, Internet hosts, and Web pages (by domains).



Sources: The Economist (population), Internet Software Consortium (Internet hosts), and Google (Web pages).