The Web of Nations

Sukwon Chung[1]
sukwonchung@stanford.edu

Dungjit Shiowattana, Pavel Dmitriev, Su Chan[2]
{dungjit,dmitriev,suchan}@yahoo-inc.com

[1]Stanford University, 450 Serra Mall, Stanford, CA, 94305, USA.
[2]Yahoo! Labs, 2821 Mission College Blvd., Santa Clara, CA 95054, USA.

ABSTRACT

In this paper, we report on a large-scale study of structural differences among the national webs. The study is based on a web-scale crawl conducted in the summer 2008. More specifically, we study two graphs derived from this crawl, the *nation graph*, with nodes corresponding to nations and edges – to links among nations, and the *host graph*, with nodes corresponding to hosts and edges – to hyperlinks among pages on the hosts. Contrary to some of the previous work [2], our results show that webs of different nations are often very different from each other, both in terms of their internal structure, and in terms of their connectivity with other nations.

Categories and Subject Descriptors

H.1.1 [**Models and Principles**]: Systems and Information Theory

General Terms

Experimentation, Measurement, Verification

Keywords: Web structure, web graph, host graph, nation graph.

1. INTRODUCTION

Graphical structure of the web has been a subject of widespread interest in research community. In their pioneering work, Broder et al. [1] found that the web on a page level looked like a bow tie, with 29% of pages in a strongly connected component (SCC), 24% being on a directed path to the SCC (IN), 24% being reachable from the SCC (OUT), and the rest of the pages not being in any of the above. In [2], Donato et al. found that, while the global web did look like a bow-tie, national webs (Italy, United Kingdom, and Indochina) looked different, with a large SCC (>50%), significant OUT (28-46%), and a very small IN. While this may have seemed like a general principle, Zhu et al. [3] showed that Chinese web was again different, having a large SCC (51%), large IN (26%), and a smaller OUT (15%).

In this paper, we report on a large-scale study of structural differences among the national webs. Our study is based on a host graph of the web obtained from a web-scale crawl, and included hosts from over 200 nations. While, similarly to the previous works, we analyze the IN/SCC/OUT structure of selected national webs, we focus more on the analysis of the *nation graph*, studying the connectivity among nations.

2. EXPERIMENTAL SETUP

Our dataset was derived from a web-scale crawl containing tens of billions of pages, conducted for the purpose of this experiment in the summer 2008. Based on this crawl we built a weighted directed host graph containing hosts as nodes and links between hosts as edges. A link from host *a* to host *b* has weight *w* iff there are *w* links from pages in *a* to pages in *b*. The graph contained over 20M nodes with over 2B edges. Then, a nation label was assigned to each host, determined either based on the top-level domain (e.g., .in, .vn), or, for hosts under non-national domains (e.g., .com, .net), based on a classification algorithm using language and link features of pages on the host. The details of the algorithm are beyond the scope of this paper[1]. Over 200 distinct nation labels were used. The nation labels allowed us isolate a subgraph of the host graph corresponding to a particular nation, and build a nation graph by collapsing each national subgraph into a single node, and adding up the edge weights. Below we present our analysis of the properties of the nation graph and the host subgraphs corresponding to selected nations. Due to space limitation, we only present a subset of our results.

3. ANALYSIS OF THE NATION GRAPH

The first property we look at is which nations tend to be linked to more than they link to other nations. Figure 1 (left) shows the ratio of inlinks to outlinks for 21 nations (nation codes are based on ISO 3166-1 standard). As one can see, India and Vietnam are two outliers with much more inlinks than outlinks, while other nations have similar ratios in the range of (0.5, 1.5).

Figure 1 (center) shows, for each nation, the average numbers for inlinks and outlinks from/to other nations per host. This graph confirms the observations made above, showing that India and Vietnam have more inlinks per host than outlinks, while Malaysia has the opposite. One can also notice that China, Korea, and Japan have the lowest number of both inlinks and outlinks to other nations per host, suggesting that most of the hosts in their webs are internally oriented. On the other end of the spectrum, Vietnam, India, and the United Kingdom are the most "social" nations, with the highest numbers of inlinks and outlinks per host.

Another interesting property is, for a particular nation, how many other nations it links to, and how many other nations link to this nation (Figure 1 (right)). Not surprisingly, the USA links to and is linked by the largest number of nations. Interestingly, Vietnam, which has the largest ratio of inlinks to outlinks, is linked by the smallest number of nations. In general, there seems to be no correlation between a nation having many inlinks from other nations, and a nations being linked to by many nations.

Figure 2 shows the proportion of inlinks and outlinks from/to other nations for the USA and Malaysia. We observed that the USA is the nation linked to and from most often, while other popular nations differ depending on the geographical location of the nation under consideration.

We also perform an analysis of the connectivity of the nation graph in terms of the sizes of the SCC, IN, and OUT components. We found that that the nation graph is well connected, with the largest SCC containing 98% nations. As we increase the threshold for edge weight required for the edge to be present in the graph, the largest SCC remains large, and IN and OUT components remain small. Figure 3 shows the central component of the graph for the threshold value of 250M. As one can see, the web seems to be centered around the USA, with the exception of United Kingdom to Germany, and Russia to Ukraine connections.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

[1] While we believe that the algorithm is quite accurate, the results described here could be affected by the mistakes made by the algorithm.

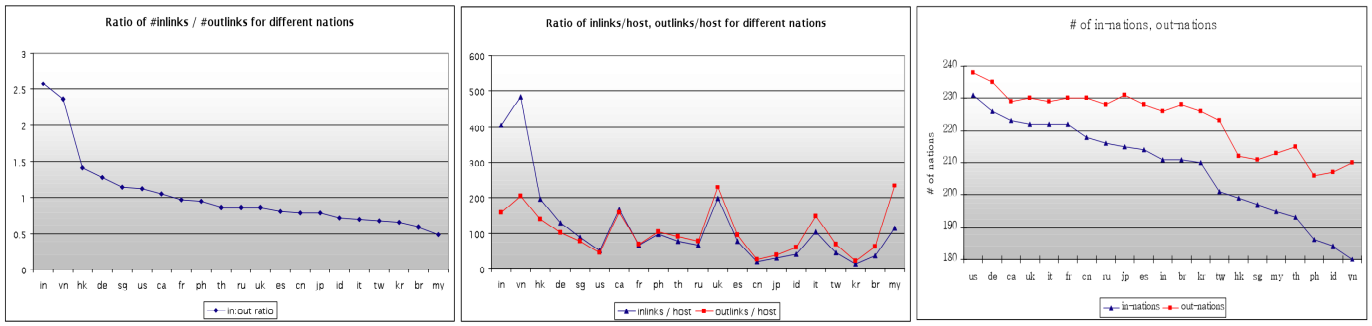

Figure 1. Inlinks/outlinks for a nation (left); inlinks/outlinks per host (center); linking and linked to nations (right)

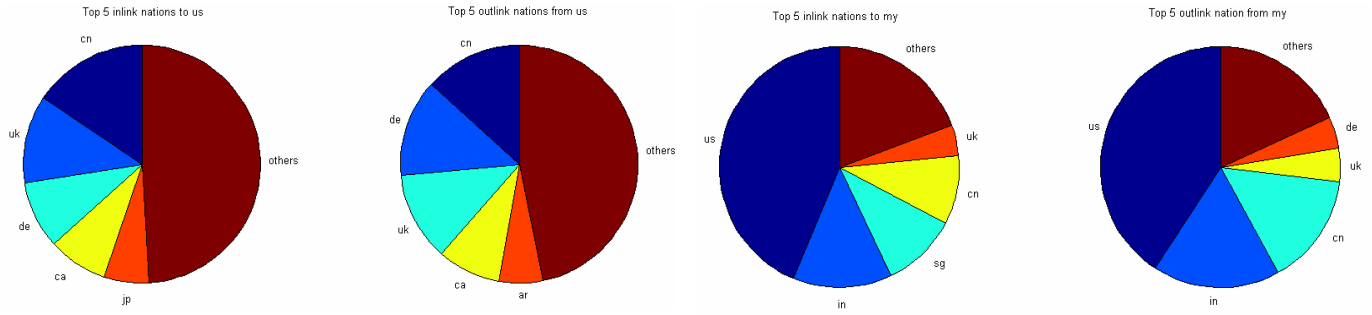

Figure 2. Proportions of inlinks and outlinks to/from other nations for Malaysia (left) and Korea (right).

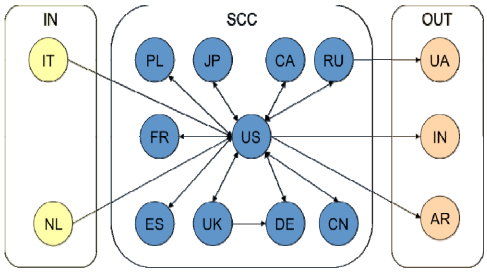

Figure 3. Nation graph SCC for 250M threshold.

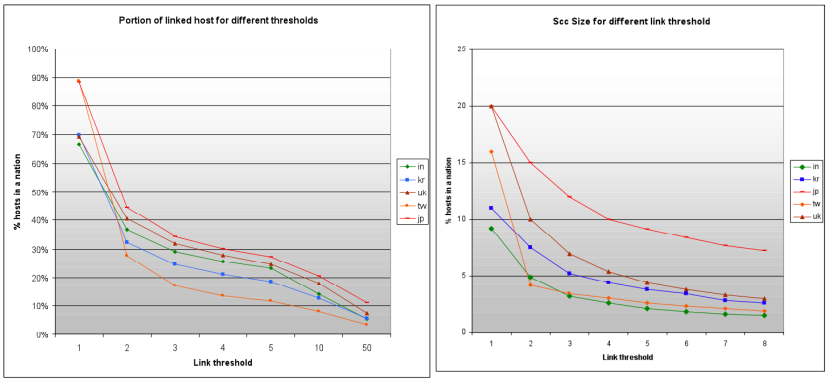

Figure 4. Percentage of hosts with external links (left); size of the largest SCC (right).

4. ANALYSIS OF THE HOST GRAPHS

Figure 4 (left) represents the percentage of hosts in a nation with at least *threshold* number of links to or from other hosts. As the graph shows, hosts within a nation are connected rather loosely, with less than 50% of the hosts having more than 1 link to/from other hosts. Taiwan has the most fragile connection structure, with almost 90% of hosts having at least 1 connection, but less than 30% having at least 2 links. Figure 4 (right) shows how the size of the largest SCC in the national web changes with the threshold. As one can see, the largest SCC contains no more than 20% hosts, and this number is dropping sharply as the threshold value increases. Japan has the strongest SCC, and Taiwan and India have the weakest SCCs.

5. CONCLUSION

In this paper, we presented a large-scale study of the structural differences among the nationals web. Our results show that webs of different nations are often very different from each other, both in terms of their internal structure, and in terms of their connectivity with other nations. We also show that, in general, host graphs of national webs are less densely connected than previous work might have suggested. We attribute this difference to bigger size and better comprehensiveness of our dataset, allowing us to discover more isolated and loosely connected hosts.

6. REFERENCES

[1] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. Graph Structure in the Web. Computer Networks, 33(1-6), pp.309-320, 2000.

[2] Donato, D., Leonardi, S., Millozzi, S., Tsaparas, P. Mining the Inner Structure of the Web Graph. 8th International Workshop on the Web and Databases (WebDB), June 16-17 2005, Baltimore, MD, USA.

[3] Zhu, J.H., Meng, T., Xie, Z., Li, G., Li, X. A Teapot Graph and its Hierarchical Structure of the Chinese Web. 17th International World Wide Web Conference, April 21-25 2008, Beijing, China.