The Small World Web

Lada A. Adamic

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
USA
`ladamic@parc.xerox.com`
WWW home page: `http://www.parc.xerox.com/iea`

Abstract. I show that the World Wide Web is a small world, in the sense that sites are highly clustered yet the path length between them is small. I also demonstrate the advantages of a search engine which makes use of the fact that pages corresponding to a particular search query can form small world networks. In a further application, the search engine uses the small-worldness of its search results to measure the connectedness between communities on the Web.

1 Introduction

Graphs found in many biological and manmade systems are "small world" networks, which are highly clustered, but the minimum distance between any two randomly chosen nodes in the graph is short. By comparison, random graphs are not clustered and have short distances, while regular lattices tend to be clustered and have long distances. Watts and Strogatz have demonstrated that a regular lattice can be transformed into a small world network by making a small fraction of the connections random [1].

Transitioning from a regular lattice to a small world topology can strongly affect the properties of the graph. For example, a small fraction of random links added to a regular lattice allows disease to spread much more rapidly across the graph. An iterated mutliplayer prisoner's dilemma game is less likely to lead to cooperation if the connections between the players form a small world network rather than a regular lattice [1]. Costs for search problems such as graph coloring have heavier tails for small world graphs as opposed to random graphs, calling for different solving strategies [2].

So far, several man made and naturally occurring networks have been identified as small world graphs. The power grid of the western US, the collaboration graph of film actors, and the neural network of the worm *Caenorhabditis elegans*, the only completely mapped neural network, have all been shown to have small world topologies. In the case of the graph of film actors, the distance between any two actors is found as follows: if the two have acted together, their minimum distance is one. If they have not costarred together, but have both costarred with the same actor, their distance is two, etc.

The concept of small worlds first arose in the context of social networks among people [3]. It has been estimated that no more than 10 or 12 links are required to go from any person to any other person on the planet via the relationship "knows," where "knows" could be defined as "can recognize and be recognized by face and name." The fact that relationships between individuals tend to form small world networks has been captured in several popular games. For example, in the game 'Six Degrees of Kevin Bacon', one attempts to find the shortest path from any actor to Kevin Bacon. Because the graph of film actors is a small world, it is difficult to find any actor with a degree of separation greater than 4 with actor Kevin Bacon. There is also the Erdos number for scientists. If a scientist has published an article with the famous Hungarian mathematician Erdos, their number is 1, if they've published with someone who's published with Erdos, their number is 2.

In this paper I show that another man-made network, the World Wide Web, has a small world topology as well. Web sites tend to be clustered, but at the same time only a few links separate any one site from any other. This topology has implications for the way users surf the Web and the ease with which they gather information. The link structure additionally provides information about the underlying relationship between people, their interests, and communities.

2 Finding Small World Properties in the Web

Watz and Strogatz define the following properties of a small world graph:

1. The clustering coefficient C is much larger than that of a random graph with the same number of vertices and average number of edges per vertex.
2. The characteristic path length L is almost as small as L for the corresponding random graph.

C is defined as follows: If a vertex v has k_v neighbors, then at most $k_v * (k_v - 1)$ directed edges can exist between them. Let C_v denote the fraction of these allowable edges that actually exist. Then C is the average over all v.

The first graph considered was the Web at the site level. Site A has a directed edge to site B, if any of the pages within A point to any page within site B. The data set used was extracted by Jim Pitkow at Xerox PARC from an Alexa crawl made approximately 1 year ago. It contains 50 million pages and 259,794 sites. Initially all links were considered to be undirected. From the 259,794 sites in the data set, the leaf nodes were removed, leaving 153,127 sites. An estimate of L was formed by averaging the paths in breadth first search trees over approximately 60,000 root nodes. 84.5% of the paths were realizable, the rest were labeled with -1. The resulting histogram is shown in Fig. 1.

L was small, a mere 3.1 hops on average between any two connected sites. C was 0.1078, compared to 2.3e-4 for a random graph with the same number of nodes and edges.

Next, directed links were considered. This was a more natural interpretation of navigation between sites, since a user cannot move in the reverse direction

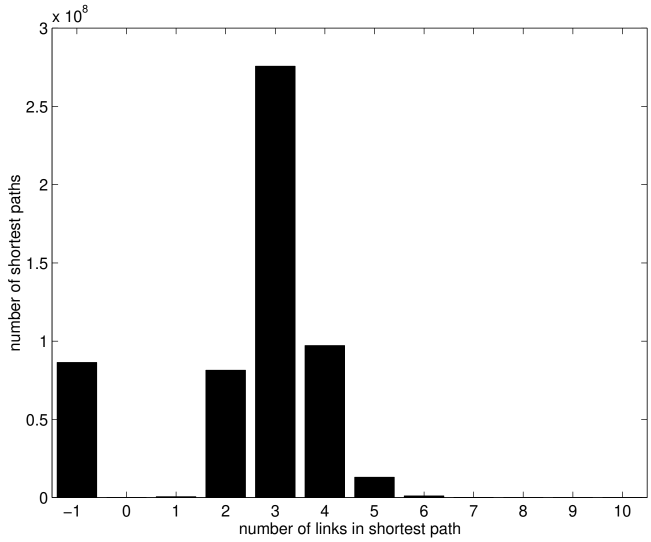

Fig. 1. Frequency of minimum path lengths between sites connected via undirected links

on links using a standard browser. The largest strongly connected component (SCC), i.e. the largest subset of nodes such that any node within it can be reached from any other node following directed links, contained 64,826 sites. In order to sample the distribution of distances between nodes, breadth first search trees were formed from a fraction of the nodes. The corresponding histogram is shown in Fig. 2.

L was slightly higher at 4.228 because the number of choices in paths is reduced when edges are no longer bi-directional. C was 0.081 compared to 1.05e-3 for a random graph with the same number of nodes and edges. In short, even though sites are highly clustered locally, one can still hop among 65,000 sites following on average only 4.2 between-site links (note that there might be additional hops within sites that are not counted in this framework). There is indeed a small world network of sites.

In order to have a more accurate comparison between the small world networks for sites, and the corresponding random graphs, the subset of *.edu* sites was considered. Because the *.edu* subset is significantly smaller, distances between every node could be computed. 3,456 of the 11,000 .edu sites formed the largest SCC. C and L were computed for a generated random graph with the same number of nodes and directed edges. A comparison between the distributions of path lengths is shown in Fig. 3.

L for the *.edu* graph was 4.062, similar to that of sites overall. This was remarkably close to L of the random graph : 4.048. At the same time C was much higher : 0.156 vs. 0.0012 for the random graph. The semi log plot in Fig. 4

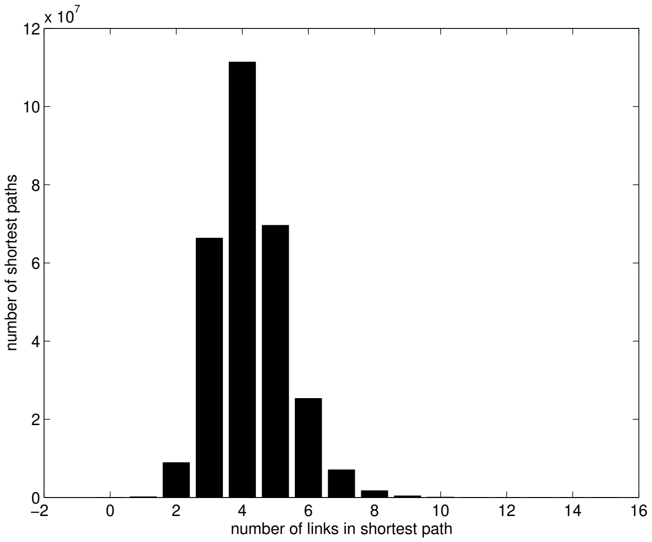

Fig. 2. Frequency of minimum path lengths between sites connected via directed links

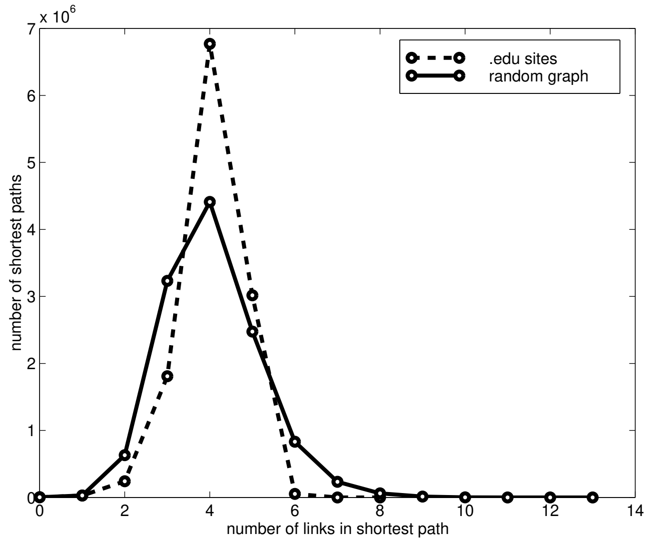

Fig. 3. Frequency of minimum path lengths between *.edu* sites compared to a random graph with the same number of nodes and edges

shows the difference in the tails of the two shortest paths histograms. While L is almost the same for both graphs, long paths (of up to 13) occur for the *.edu* site graph. For the corresponding random graph the maximum path is 8 and long paths are unlikely. While the average shortest path was almost identical, the small world network distinguishes itself by having a few unusually longs paths as well as a much larger clustering coefficient.

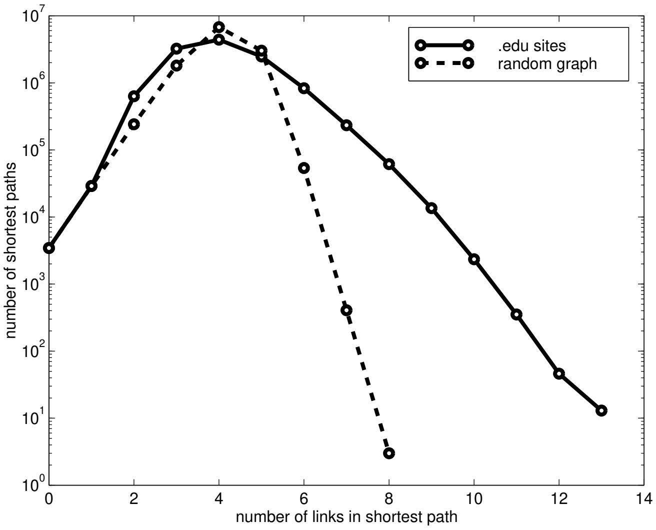

Fig. 4. Semilog plot of the frequency of minimum path lengths within the *.edu* site graph compared to a random graph with the same number of nodes and edges.

In summary, the largest SCCs of both sites in general and the subset of *.edu* sites are small world networks with small average minimum distances.

3 A Smarter Search Engine Using Small World Properties

3.1 Introduction

The above analysis showed that at the site level the Web exhibits structure while staying interconnected. One would expect a similar behavior at the page level. Related documents tend to link to one another while containing shortcut links to documents with different content. This small world link structure of pages can be used to return more relevant documents to a query.

Links are interpreted as a citations of one document by another. Citations have been used to evaluate the impact of journals and authors [9][10]. Here they are used to identify quality web pages. Starting from the assumption that a good Web document references other good documents of similar content, one would expect that there exist groups of pages of similar content which refer to one another. The quality of these pages is guaranteed by the recommendations implicit in the links among them. Such groups can be extracted from results matching a particular query. Within each group there are documents which are good "centers", that is, the distance from them to any other document within the group is on average a minimum. Centers tend to be index pages, and hence constitute good starting points for exploration of related query results.

An application of these ideas was built around webbase, a repository of Web pages crawled by Google (`http://www.google.com`) in the first half on 1998. For any given search string, webbase returns queries ranked by a combination of their text match score and their PageRank[11], which is based on the links to the document. Webbase also provides link information for each page.

3.2 Outline of the Application

1. Query webbase for docids corresponding to a particular search string.
2. Identify all SCCs within the search results.
3. Identify the largest SCC.
4. Calculate L from each node in the largest SCC to find the best center.
5. Form a minimum spanning tree via breadth first search from the best center (a graphical interface could guide the user down the tree).
6. Compute C for the largest SCC.

3.3 Observations

SCCs usually contain pages belonging to the same site, because pages within a site are more likely to be linked to one another than pages across sites. A preference should be given to SCCs spanning the most sites, because links across sites are stronger recommendations than links within a single site. SCCs containing the same number of sites are ordered by the number of documents they contain. A large number of interconnected documents implies a degree of organization and good coverage of the query terms. Rather than presenting a list of documents that contains many sequential entries from the same site, the search engine can present just the center from each SCC. By sorting the centers by the size of their SCCs, one can present the user with the maximum span of the search space with the minimum number of entries. Given a good starting point, the users can explore the SCCs on their own.

Informal observation suggests that pages which do not belong to any large SCC tend to focus on either a narrow topic, don't have many outlinks, or don't have many pages referencing them (which implies that they are probably not worth reading). When centers are sorted by the size of their SCCs, these documents will be listed last. One further observes that the SCCs that span several sites tend to contain the main relevant pages, and are rich in "hubs", or pages that contain links to many good pages. The algorithm will present these at the top of the list.

3.4 An Example

A good example of how extracting SCCs and finding good centers can expedite a search, is in the query for "java". Webbase returned 311 documents. The largest SCC contained 144 pages ranging over 28 sites, with the vast majority of pages belonging to `java.sun.com/`.

If we look at the top 10 centers (see Table 1) among the 144 pages in the largest SCC, we see that they are compilations of links to java resources on the Web. Such pages are obvious good starting points. Since they are contained in the SCC, they must be linked to by at least one other page, which means that this list has been evaluated and deemed useful by at least one other source. Pages which are compilations of links tend to be subdivided into topics, and have brief human evaluations or summaries of the pages linked to. If a search engine is able to present the user with such man made sources of links, it could potentially save itself the trouble of ranking, clustering, or otherwise evaluating documents matching the search string.

It is interesting to note that this procedure might not allow the search engine to return the best resource immediately, but with high probability the best resource is only one step away. For example, java.sun.com is not in the top 5 centers, but all 5 centers link to it. One of the top five documents returned by webbase, `www.gamelan.com` does not appear in the SCC. It has many back links, but no forward links, and hence is disconnected from the other pages. Still, 4 of the top 5 centers listed reference Gamelan, so that, again, this important site is just a click away from the center. What if, on the other hand, one were interested in just the best single page on java? Then one could look for the page that has a minimum average distance to it from any other page in the SCC. As Table 2 shows, `java.sun.com` comes out on top, as do other good "authorities" - pages which are good single sources of information. Rather than revealing good starting points to explore the topic, this approach brings the user directly to the most authoritative pages on the subject. Just as once it was said that all roads lead to Rome, so it can be said that all links lead to the most important pages. In summary, in the "java" search string example, the largest SCC is a group of high quality resources spanning several sites. Pages with lowest average distance to any other page within the largest SCC are good hubs, i.e. pages with many links to good sources. Pages with lowest average distance to them from any other page in the largest SCC tend to be good sources of information, or "authorities" on java.

Future work will include a comparison with the Kleinberg algorithm [8] which also finds "hubs" and "authorities" using the link structure of the Web.

4 Marketing Applications

The small world topology can facilitate navigation and extraction of useful information, but it can also give clues to the structure of communities from the link structure of the documents they create. The Web today represents a wide

Table 1. Top 10 centers for the LSCC for the search string "java"

Av. Min. Dist.	URL & Title
2.47222	`http://www.infospheres.caltech.edu/resources/java.html` Infospheres - Java Resources
2.48611	`http://www.apl.jhu.edu/ hall/java/` Java Programming Resources: Java, Java and More Java.
2.70138	`http://sunsite.unc.edu/javafaq/links.html` Java Links
2.73611	`http://www.cat.syr.edu/3Si/Java/Links.html` 3Si - Java Resources
2.77777	`http://www.december.com/works/java/info.html` Presenting Java: Information Sources
2.79861	`http://www.javaworld.com/javaworld/common/jw-jumps.html` JavaWorld - Java Jumps
2.93055	`http://java.sun.com/aboutJava/jug/` Java(TM) User's Groups Info Page
3.01388	`http://javaboutique.internet.com/javafaqs.html` The Java(TM) Boutique: Java FAQs
3.04166	`http://sunsite.unc.edu/javafaq/` Cafe au Lait Java FAQs, News, and Resources
3.14583	`http://java.sun.com/` The Source for Java(TM) Technology

Table 2. Top 10 attractors of the largest SCC for the search string "java"

Av.Min.Dist	URL & Title
1.90972	`http://java.sun.com/` The Source for Java(TM) Technology
2.29861	`http://java.sun.com/products/` Products & APIs
2.3125	`http://java.sun.com/applets/` Applets
2.33333	`http://java.sun.com/nav/used/` Java(TM) Technology in the Real World
2.34722	`http://java.sun.com/docs/` Documentation
2.35416	`http://java.sun.com/nav/developer/` For Developers
2.53472	`http://java.sun.com:81/` Java Home Page
2.63888	`http://java.sun.com/sfaq/` Frequently Asked Questions
2.66666	`http://java.sun.com/javaone/` JavaOne Home
2.74306	`http://java.sun.com/products/activator/` Java Development

range of human interests. Some sites are devoted entirely to a single interest or cause. Others, such as Yahoo, have clubs or chat rooms where people can meet and share their ideas on particular topics. Many people document their interests and affiliations in their personal home pages. Therefore exploring the link structure of documents which belong to a particular topic could reveal the underlying relationship between people and organizations.

To see what insight one could gain from identifying strongly connected components and average shortest paths, three search strings were issued to the search engine application outlined above: "abortion - pro choice"[1], "abortion - pro life"[2], and "UFO"[3].

Although the pro choice results contained several sites devoted entirely to the issue, such as `www.choice.org`, `www.cais.com`, `www.abortion.com`, and `www.prochoice.com`, these sites did not appear to be linked to one another (i.e. there was no strongly connected component containing pages from more than one site). In fact, the largest strongly connected component was a group of pro life pages which had mentioned "pro choice" in their content.

On the other hand, the pro life query results had a pro life strongly connected component of 41 pages, which spanned 16 sites. One could conclude that pro lifers not only have a stronger Web presence (804 vs. 645 documents returned for the two queries), but that the pro life community is more tightly knit, and possibly better organized.

The results of the UFO query contained a largest connected component of 95 pages, spanning 21 sites. Apparently there is a lot of interest in UFOs and UFO enthusiasts are interested in other's sightings and speculations.

The largest strongly connected components for all three queries had a high clustering coefficient and a small average shortest path, showing that groups of people with common interests are linked to one another via a small world network on the Web.

How does this tie into marketing? Suppose one were interested in informing others of upcoming legislation regarding abortion. For example, a while back one could have either opposed or supported the partial birth abortion ban bill and wanted to start a red ribbon campaign. A ribbon placed on any site acts as a link to the main campaign site. The main site provides information about the campaign and allows others to download and include ribbons in their own sites. One could place one red ribbon in support of the bill in the middle of the pro life strongly connected sites and expect your ribbon to find its way to other pro life sites. On the other hand, if one wanted to start a black ribbon campaign in opposition to the bill, one would have to drop a black ribbon at several pro choice sites, because one would not expect the ribbon to propagate on its own. In general, one could reach a large community of people by placing an ad on a central page of an SCC. If the community is represented on the Web by many

---

[1] Data can be viewed at `http://www.stanford.edu/~ladamic/data/pccenters.htm`.
[2] Data can be viewed at `http://www.stanford.edu/~ladamic/data/plcenters.htm`.
[3] Data can be viewed at `http://www.stanford.edu/~ladamic/data/ufocenters.htm`.

small SCCs, the advertiser would need to place ads in many SCCs, in order to ensure reaching as much of the target audience as possible.

5 Conclusions

I have shown that the largest strongly connected component of the graph of sites on the Web is a small world. The graph of all sites and of the .edu subset has an average minimum distance between nodes that is close to that of a random graph with the same number of nodes and edges. At the same time both sets of sites are highly clustered. These two properties make the Web a small world, at least at the site level. I have developed a prototype of a search engine application that can take advantage of the small world networks present in documents corresponding to particular queries. In the example of the "java" search string, the application could present theuser with documents which are good starting points for exploring, a maximum number of quality sites within a minimum distance or it could cut to the chase and return the highest quality documents directly. Finally, I have used the application to draw inferences about the connectedness of several communities of people that are represented on the Web and how it could influence advertising strategy.

6 Acknowledgement

I thank Jim Pitkow for use of his data and helpful discussions.

References

1. Watts, D., Strogatz, S.: Collective dynamics of 'small world' networks. Nature **393** (1998)
2. Walsh, T.: Search in a Small world Report APES-07-1998
3. Pool, I., Kochen, M.: Contacts and influence. Social Networks **1** (1978) pp.5-51
4. Google
   `http://www.google.com`
5. Matthews, R. Six degrees of separation New Scientist 6 June, 1998
6. Oracle of Bacon at Virginia
   `http://www.cs.virginia.edu/oracle`
7. Erdos Numbers
   `http://www.acs.oakland.edu/ grossman/erdoshp.html`
8. Cormen, T., Leiserson, C, Rivest, R.: Introduction to Algorithms, MIT Press, 1990.
9. Garfield, E.: Citation Analysis as a Tool in Journal Evaluation" Science **178** (1972), pp.471-479
10. Garfield, E.: The Impact Factor Current Contents, June 20, 1994.
11. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine
    `http://google.stanford.edu/long321.htm`
12. Kleinberg, J: Authorative Sources in a Hyperlinked Environment IBM Research Report RJ 10076, May 1997.