Graph Theory, Small Worlds, and the WEB: selection of some topics for Spring School 2001

Bruno Codenotti*

April 5, 2001

1 Introduction

The small world phenomenon formalizes and generalizes the anecdotal notion of two strangers who, after unexpectedly discovering that they have a common acquaintance, exclaim "it's a small world!"

We will illustrate this phenomenon, and analyze its potential impact on computational issues related to WEB crawling and searching. We will start our review from folklore examples, and from experiments done in the sixties in the social sciences (Milgram, Psychology Today, 1967), which gave rise to the (by now popular) expression "six degrees of separation".

We will then see how this pervasive notion later got the attention of the exact sciences (Watts and Strogatz, Nature 1998) and finally of computer science. We will especially focus on

- the background in graph theory which must guide our understanding of the main issues concerning small worlds,
- the latest developments in the context of "decentralized" algorithms (Kleinberg, Proc. 32nd ACM STOC 2000, and Nature, 2000), and
- applications to the WEB (Albert et al, Nature, 1999).

*Istituto di Matematica Computazionale del CNR, Area della Ricerca di S.Cataldo, 56010-Ghezzano, Pisa (Italy). e-mail: codenotti@imc.pi.cnr.it.

Most of the papers cited here can be easily obtained on-line.
The students should

- first get the papers on random graphs, especially those focusing on the diameter of random graphs (see the list of papers at the end of Section 4),
- then familiarize with the notion of small world (see Section 3, and in particular http://aries.mos.org/internet/essay.html and

  http://www.discover.com/dec_issue/smallworld.html, for an easy and popularized presentation), and
- should read at least the papers written by Jon Kleinberg, on the possibility of finding decentralized algorithms within small world topologies, which can be obtained on-line (see

  http://www.cs.cornell.edu/home/kleinber/swn.ps, and also http://cornell-magazine.cornell.edu/Archive/JulyAugust98/JulyWorld.html for an introduction to nonspecialists).
- Finally the reader should look at the literature where the WEB is viewed as a graph (see Section 5); in particular should download the paper "Diameter of the World-Wide Web" by Albert, Jeong and Barabási from the site http://xxx.lanl.gov/abs/cond-mat/9907038. Here is what the abstract of the paper says: "Despite its increasing role in communication, the world wide web remains the least controlled medium: any individual or institution can create websites with unrestricted number of documents and links. While great efforts are made to map and characterize the Internet's infrastructure, little is known about the topology of the web. Here we take a first step to fill this gap: we use local connectivity measurements to construct a topological model of the world wide web, allowing us to explore and characterize its large scale properties."

2 Small worlds: Milgram's experiment

Informally, we talk of a social network, whenever we have a (typically large) number of people, each corresponding to a node in a network, and we represent the presence of a relationship (which could be acquaintanceship, friendship, etc.) between any two people by an edge connecting the two nodes. We say that a social network exhibits the small-world phenomenon if, roughly speaking, any two individuals in the network are likely to be connected through a short sequence of intermediate acquaintances.

This has long been the subject of anecdotal observation and folklore; often we meet a stranger and discover that we have an acquaintance in common. It has since grown into a significant area of study in the social sciences, in large part through a series of striking experiments conducted by Stanley Milgram and his co-workers in the 1960's.

Recent work has suggested that the phenomenon is pervasive in networks arising in nature and technology, and a fundamental ingredient in the structural evolution of the World Wide Web.

Milgram's basic *small-world experiment* remains one of the most compelling ways to think about the problem. The goal of the experiment was to find short chains of acquaintances linking pairs of people in the United States who did not know one another. In a typical instance of the experiment, *a source person in Nebraska would be given a letter to deliver to a target person in Massachusetts. The source would initially be told basic information about the target, including his address and occupation; the source would then be instructed to send the letter to someone she knew on a first-name basis in an effort to transmit the letter to the target as efficaciously as possible. Anyone subsequently receiving the letter would be given the same instructions, and the chain of communication would continue until the target was reached.*

Over many trials, the average number of intermediate steps in a successful chain was found to lie between five and six, a quantity that has since entered popular culture as the "six degrees of separation" principle. (From http://www.cs.cornell.edu/home/kleinber/swn.d/swn.html.)

3 List of WEB sites with papers and references to small worlds

- http://elegans.swmed.edu This site contains information on a tiny worm (C.elegans) whose 302 (!) neurons are connected in order to form a small world.

- http://www.santafe.edu/sfi/publications/Bulletins/

bulletinFall99/workInProgress/smallWorld.html

This site contains an article on The Kevin Bacon Game (Santa Fe Inst. Bulletin 14(2)). Here is an excerpt from the article:

“This is how the game goes: Think of an actor or actress. If they have ever been in a film with Kevin Bacon, then they have a ”Bacon Number” of one. If they have never been in a film with Kevin Bacon but have been in a film with somebody else who has, then they have a Bacon Number of two, and so on. The claim is that no one who has been in an American film, ever has a Bacon Number of greater than four.”

- http://www.cs.virginia.edu/oracle/ The Oracle of Bacon at Virginia: on-line tools to actually try Kevin Bacon game...
- www.sixdegrees.com This is a site which actually implements small worlds, by finding ways to link people on the WEB, etc.
- http://www.cs.washington.edu/homes/kautz/referralweb/index.html

  Web site (with demo) for building graphs of scientific collaboration.
- http://www.amsci.org/amsci/issues/Comsci00/compsci2000-01.html Here you can find the paper: Graph Theory in Practice: Part I, by Brian Hayes (January-February, Volume 88, No. 1 Computing Science).
- http://www.amsci.org/amsci/issues/Comsci00/compsci2000-03.html Here you can find the paper: Graph Theory in Practice: Part II, by Brian Hayes (March-April, Volume 88, No. 2 Computing Science).
- http://aries.mos.org/internet/essay.html Contains the article on “The Internet, The Small World, and The Nature of Distance”, by Duncan Watts.
- http://www.discover.com/dec_issue/smallworld.html Another on-line introduction to small worlds, for the general audience.

4 Background in graph theory

The content of this section has been taken from
http://www.amsci.org/amsci/issues/Comsci00/compsci2000-03.html.

Erdös and Rényi proved a number of interesting results about random graphs. Many proofs concerns properties that hold for “almost every” random graph; this sounds like a strangely vague manner of speaking for mathematical discourse, but it has a precise meaning. Saying that almost every random graph has some property Q means that as the size of the graph n goes to infinity, the probability of Q approaches 1. For example, Erdös and Rényi showed that if the edge probability p is greater than a certain threshold, then almost every random graph is connected. This doesn’t mean you can’t construct disconnected graphs if you set out to do so, but the random process has no chance of producing them when n approaches infinity.

As a model of small-world networks, the Erdös-Rényi random graph has some strengths. It can be made as dense or as sparse as necessary just by adjusting the edge probability p. And the diameter tends to be small.

On the other hand, the difference with respect to small world graphs is that Erdös-Rényi graphs show no tendency to form clusters, since the edges are placed independently, and neighbors of neighbors are no more likely to be linked than any other randomly chosen vertices. This contrasts with the fact that, in a small world, clusters are naturally formed (informally, it’s likely that friends of friends are themselves friends).

The following papers are relevant to the issues of graph diameter which will play a crucial role later on:

- Erdös and Rényi. 1960. On the evolution of random graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5:17-61.
- Aiello, William, Fan Chung and Linyuan Lu. 1999. A random graph model for massive graphs. Proceedings of the 32d Annual ACM Symposium on the Theory of Computing. http://math.ucsd.edu/ fan/random.ps
- Abello, J., P. M. Pardalos and M. G. C. Resende. 1999. On maximum clique problems in very large graphs. In External Memory Algorithms (J. Abello and J. Vitter, eds.), AMS-DIMACS Series on Discrete Mathematics and Theoretical Computer Science, Vol. 50.

- Bollobás, B., and F. R. K. Chung. 1988. The diameter of a cycle plus a random matching. SIAM Journal on Discrete Mathematics 1:328-333.
- Chung, F. R. K., and M. R. Garey. 1984. Diameter bounds for altered graphs. Journal of Graph Theory 8:511-534.

5 Background: graph theory and the WEB

The content of this section has been taken from
http://www.mills.edu/ACAD_INFO/MCS/CS/S00MCS125/
Practical.Graph.Theory/math125.html.

Arguably the most widely known and broadly accessed graph being studied today is the World Wide Web. Although not popularly thought of as a graph, one can consider the different Web pages to be vertices and the links between them to be edges, making the Web a graph.

Granted, the Web is an enormous graph; it has some 800 million vertices. It has only been recently that graph theory has been thought to be useful in studying such a tremendous structure. Due to the size of the graph of the Web, studying it requires that one move sections of it in and out of memory and disk storage, a procedure that destroys the efficiency models of most algorithms run on these sections. In order to make the study of such large graphs more accessible, they are reduced from directed multigraphs to undirected simple graphs.

The Web is one of a family of large "small-world" graphs. Anyone familiar with the term "six degrees of separation" already understands the main characteristics of a small-world graph.

These graphs are generally sparse. Any undirected, simple graph with n vertices can have at most $\frac{n^2}{2}$ edges. In large small-world graphs, the tendency of the number of edges is more along the lines of n edges. Further, small-world graphs are often clustered graphs, that is, the edges of the graph do not have uniform distribution. Instead, they tend to form knots. Lastly, the diameter, i.e., the length of the most direct route between the most distant vertices, is usually small.

Studies have estimated that the Web, being a small world graph, has a diameter of 19 (see [2], and the online version mentioned at the end of this section). This means that to get from one randomly selected Web page to a page one desires requires an average of 19 clicks.

Graphs like the *call graph*[1], the Oracle of Bacon Graph, and the Erdös number graph are all examples of small-world graphs. Only the call graph, with about 54 million vertices and 170 million edges generated in a 20 day graph, is comparative to the size of the Web.

In order to better understand the structure of the graph of the Web, scientists have attempted to find algorithms to construct graphs that share the same properties. The problem there lies in the fact that the Web has grown very organically and thus is neither random nor regular. After several attempts by various scientists to create a graph that would grow to have a similar structure as the Web, researchers at Notre Dame decided to create an amalgam of a regular ring lattice graph and a random graph to mimic the properties of the Web.

See the following papers:

- Albert, Reka, Hawoong Jeong and Albert-László Barabási. 1999. Diameter of the World-Wide Web. Nature 401:130-131. http://xxx.lanl.gov/abs/cond-mat/9907038
- Adamic, Lada. "Testing the Small-World Hypothesis on the Web" (1999). A ps file with the paper can be downloaded from:

  http://www.parc.xerox.com/istl/groups/iea/www/smallworld.html

  Here is the abstract of this paper: "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] A good example of a really big graph comes from telephone billing records. The vertices of this =93call graph=94 are telephone numbers, and the edges are calls made from one number to another. James M. Abello of the AT&T Shannon Laboratories in Florham Park, New Jersey, has studied the evolution of the graph as calls accumulate over a period of days. In one 20-day period the graph grew to have 290 million vertices and 4 billion edges.(From "Graph Theory in Practice: Part I", by Brian Hayes.)

6 Watts and Strogatz results

Most of the early work on the small world issue was based on versions of the following explanation: random networks have low diameter. (See for example the book of surveys edited by Kochen [13].) That is, if every individual in the United States were to have a small number of acquaintances selected uniformly at random from the population – and if acquaintanceship were symmetric – then two random individuals would be linked by a short chain with high probability. Even this early work recognized the limitations of a uniform random model; if A and B are two individuals with a common friend, it is much more likely that they themselves are friends[2]. But at the same time, a network of acquaintanceships that is too "clustered" will not have the low diameter that Milgram's experiments indicated.

Recently, Watts and Strogatz proposed a model (WS model) for the small-world phenomenon based on a class of random networks that interpolates between these two extremes, in which the edges of the network are divided into "local" and "long-range" contacts. The paradigmatic example they studied was a "re-wired ring lattice," constructed roughly as follows. One starts with a set V of n points spaced uniformly on a circle, and joins each point by an edge to each of its k nearest neighbors, for a small constant k. These are the "local contacts" in the network. One then introduces a small number of edges in which the endpoints are chosen uniformly at random from V – the "long-range contacts". Watts and Strogatz argued that such a model captures two crucial parameters of social networks: there is a simple underlying structure that explains the presence of most edges, but a few edges are produced by a random process that does not respect this structure. Their networks thus have low diameter (like uniform random networks), but also have the property that many of the neighbors of a node uare themselves neighbors (unlike uniform random networks). They showed that a number of naturally arising networks exhibit this pair of properties (including the connections among neurons in the nematode species C.elegans, and the power grid of the Western U.S.); and their approach has been applied to the analysis of the hyperlink graph of the World Wide Web as well.

Networks that are formed from a superposition of a "structured subgraph" and a "random subgraph" have been investigated in the area of probabilistic

[2] The reader should compare this with the discussion of Section 4.

combinatorics. In a fundamental instance of such an approach, Bollobás and Chung [4] gave bounds on the diameter of the random graph obtained by adding a random matching to the nodes of a cycle. (See [10, 11, 12], and also the site http://aries.mos.org/internet/essay.html, where an online paper by Watts is available.)

7 Kleinberg results

Kleinberg comments on the success of Milgram's experiment, saying that it "suggests a source of latent navigational cues embedded in the underlying social network, by which a message could implicitly be guided quickly from source to target". (See [5, 6, 7], and the on-line version mentioned in the Introduction.)

And also: "Existing models are insufficient to explain the striking algorithmic component of Milgram=92s original findings: that individuals using local information are collectively very effective at actually constructing short paths between two points in a social network".

As an example of the fact that existing models are insufficient to explain the success of decentralized algorithms, he finds a counterexample for WS model. Kleinberg introduces a family of models generalizing WS. He proves that

(1) for one of these models there is a decentralized algorithm capable of finding short paths w.h.p., and

(2) there exists only a unique model within the family such that (1) holds.

The setup for the decentralized algorithm reflects Milgram=92s experiment and is as follows.

The model is given by a lattice (grid) with the addition of a few long range connections chosen randomly.

The basic ingredients are the following:

- Lattice distance between $u = 3D(i, j)$ and $v = 3D(k, l)$:

$$d(u, v) = 3D|k - i| + |l - j|\,.$$

- Universal constant $p \geq 1$ s.t. each node has a directed edge to every other node at distance at most p (local contacts). For the 2-dimensional grid, $p = 3D1$.
- Universal constants q , r $\geq$ 0 s.t. each node u has directed edges to q other nodes (long range contacts), where the i-th edge from u has endpoint v with probability proportional to $[d(u, v)]^{-r}$.

The message holder at any given step has knowledge of:

a. underlying grid structure;

b. location of the target on the grid;

c. location of long range contacts of all the nodes that have been involved with the msg.

REMARKS:

- (c) is used only for lower bounds; upper bounds use only (a) and (b).
- Constraining the algorithm to use only local information is crucial; otherwise it is easy to find shortest path between any two nodes.

Decentralized Algorithm: Description and Analysis

The algorithm is extremely simple:

At each step, the current msg holder chooses a contact that is as close as possible to the target (in terms of lattice distance).

This actually matches how short chains are found in real life (as in Milgram=92s experiment).

Given node u , consider sets of nodes $A_0, A_1, A_2, \ldots, A_{\log n}$, where A_j is the set of nodes at distance between 2^j and 2^{j+1} from u.

It turns out that $r = 3D2$ is the unique exponent at which a node's long-range contacts are nearly uniformly distributed over all "distance scales". (The student should now really get Kleinberg's paper and study the proof of this, which is not difficult; in the following we just sketch the idea of the algorithm's analysis for $r = 3D2$).

For $r = 3D2$, the algorithm is divided into phases, numbered $\log n, \log n - 1, \dots, j, \dots, 2, 1, 0$. In phase j, the distance from the target is between 2^j and 2^{j+1}. Phase j takes time $\log n$: the expected time before the current message holder has a long-range contact at distance less than 2^j from the target is $\log n$. Overall there are $\log n + 1$ phases, so that total time is roughly $\log^2 n$.

The main question underlying the issue under investigation is:
How does the structure of the network affect the ability of a decentralized algorithm to construct a short path?
Kleiberg's findings show that the answer is:

- for $r = 3D0$: with h.p. there exist short paths, but there is no way= a decentralized algorithm can find them (see Kleinberg's paper, where it is proved that the expected delivery time is about $n^{2/3}$).
- as r increases: decentralized algorithms can take more and more advantage of the geographic structure of long-range contacts; at the same time long-range contacts are less and less useful in moving messages a large distance.

It turns out that the tradeoff best exploited algorithmically is for $r = 3D2$, which leads to an expected delivery time of about $log^2 n$.
Instead, when long range contacts are formed independently of the geometry of the grid, short chains will exist, but the nodes, operating at a local level, will not be able to find them. For $r = 3D2$, short chains will still exist, and the nodes, with local knowledge, will be able to construct them.
In the following, we propose some general observations, which should guide the students in their reading of the original papers.

- Correlation between local structure and long range connections provides fundamental cues for finding paths through small-world networks.
- When the correlation is near a critical threshold, the structure of long-range connections forms a sort of gradient that allows individuals to guide a message efficiently towards the target.
- Below the critical value the cues disappear.

References

[1] Adamic, The small world WEB (Proc. Europ. Conf. on Dig. Libr.,1999)

[2] Albert et al., The diameter of the WWW (Nature 401, 1999)

[3] Berners-Lee, The fractal nature of the Web (Axioms of WEB Architectures, 2000)

[4] Bollobás, B., and F. R. K. Chung. 1988. The diameter of a cycle plus a random matching. SIAM Journal on Discrete Mathematics 1:328-333.

[5] J. Kleinberg. Navigation in a Small World. Nature 406(2000)

[6] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. Cornell Computer Science Technical Report 99-1776, October 1999. (This is an extended version of the Nature paper.)

[7] J. Kleinberg , The small-world phenomenon: an algorithmic perspective, Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000 (conference version).

[8] Milgram, S. (1967). The small-world problem. Psychology To= day, 1(1), 62-67

[9] Milgram, Stanley. 1977. The small world problem. In “The Individual in a Social World: Essays and Experiments”, pp. 281-295. Reading, Mass.: Addison-Wesley.

[10] Duncan J. Watts, Small Worlds: The Dynamics of Networks be= tween Order and Randomness, Princeton Univ. Press, 1999.

[11] Watts, D. J. Networks, dynamics and the small-world phenom= enon. American Journal of Sociology 105(2):493-527 (1999 a.).

[12] Watts D. J. and Strogatz S. H. Collective dynamics of 'small-world' networks. Nature 393, 440-442 (1998).

[13] M. Kochen, Ed., The Small World (Ablex, Norwood, 1989).

[14] Hayes, Brain, 2000, Graph-Theory in Practice, Part I, Amer= ican Scientist 88:1, 9-13.

[15] Hayes, Brian, 2000, Graph-Theory in Practice, Part II, American Scientist 88:2,104-109.

[16] Erdös, P., and A. Rényi. 1960. On the evolution of ran= dom graphs. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 5:17-61.

[17] Rapoport, Anatol. 1957. Contribution to the theory of rando= m and biased nets. Bulletin of Mathematical Biology 19:257-277.

[18] Saulnier, Beth. 1998. Small world. Cornell Magazine 101(1):24-29. http://cornell-magazine.cornell.edu/Archive/JulyAugust98/JulyWorld.html