Discovery and Existence of Communities in the World Wide Web

Antonis Sidiropoulos
Aristotle University of Thessaloniki, Greece

Dimitrios Katsaros
Aristotle University of Thessaloniki, Greece
University of Thessaly, Volos, Greece

Yannis Manolopoulos
Aristotle University of Thessaloniki, Greece

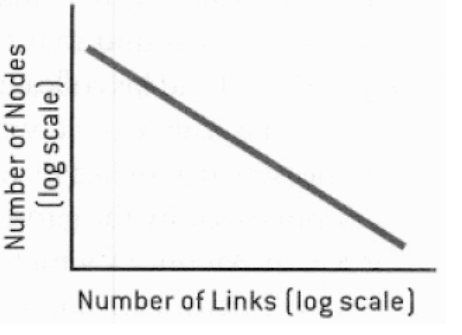
INTRODUCTION

The World Wide Web, or simply Web, is a characteristic example of a social network (Newman, 2003; Wasserman & Faust, 1994). Other examples of social networks include the food web network, scientific collaboration networks, sexual relationships networks, metabolic networks, and air transportation networks. Socials networks are usually abstracted as graphs, comprised by vertices, edges (directed or not), and in some cases, with weights on these edges. Social network theory is concerned with properties related to connectivity (degree, structure, centrality), distances (diameter, shortest paths), "resilience" (geodesic edges or vertices, articulation vertices) of these graphs, models of network growth. Social networks have been studied long before the conception of the Web. Pioneering works for the characterization of the Web as a social network and for the study of its basic properties are due to the work of Barabasi and its colleagues (Albert, Jeong & Barabasi, 1999). Later, several studies investigated other aspects like its growth (Bianconi & Barabasi, 2001; Menczer, 2004; Pennock, Flake, Lawrence, Glover, & Giles, 2002; Watts & Strogatz, 1998), its "small-world" nature in that pages can reach other pages with only a small number of links, and its scale-free nature (Adamic & Huberman, 2000; Barabasi & Albert, 1999; Barabasi & Bonabeau, 2003) (i.e., a feature implying that it is dominated by a relatively small number of Web pages that are connected to many others; these pages are called hubs and have a seemingly unlimited number of hyperlinks). Thus, the distribution of Web page linkages follows a power law in that most nodes have just a few hyperlinks and some have a tremendous number of links In that sense, the system has no "scale" (see Figure 1).

One of the most intriguing features of the Web, and of other social networks as well, is its self-organization behavior, which is usually faced through the existence of communities. Groups of vertices that have high density of edges within them, with a lower density of edges between groups is a frequent, informal definition of a community. The notion of community is very useful from a practical perspective because it can be used to improve the effectiveness of search engines (Flake, Lawrence, Giles, & Coetzee, 2002b; Radicchi, Castellano, Cecconi, Loreto, & Parisi, 2004), for purposes of prefetching (Sidiropoulos, et al., 2007), bibliographic citation ranking (Sidiropoulos, Katsaros, & Manolopoulos, 2006), spam detection (Gibson, Kumar, & Tomkins, 2005), etc.

The present article aims at presenting the notion of Web communities, providing quantitative definitions for them, and highlighting the different techniques that have been developed to discover such structures in the Web. It by no means intends to serve as a survey, but as a comprehensive introduction into the specific area.

Figure 1. Typical linkage distribution for scale-free networks

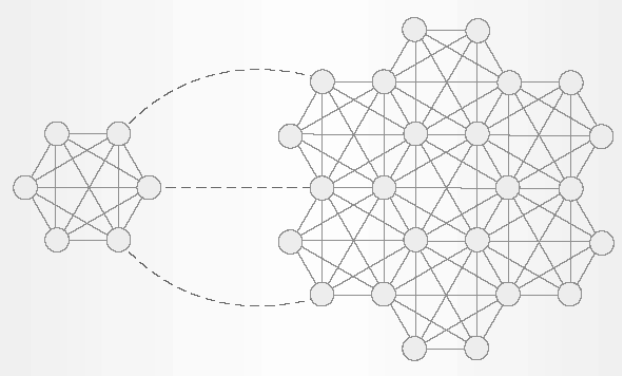


THE NOTION OF A COMMUNITY

The notion of a Web community is not very strict; it is generally described as a substructure (subset of vertices, for example, subsets of Web pages) of a Web graph with dense linkage (hyperlinks) between the members of the community and sparse density outside the community. Equipped with such a description, it is very easy to visually identify two communities in the graph depicted in Figure 2. The existence of communities in the Web was first reported in Gibson, Kleinberg, and Raghavan (1998). The aforementioned qualitative definition though is not adequate when trying to devise algorithms for the determination of communities in Web graphs. Thus, we need sharper, quantitative definitions for the communities.

In order to provide such a quantitative definition, we need to introduce some "quantities." The basic quantity to consider is di, the degree of a generic node i of the considered graph G (representing the examined network), which, in terms of its adjacency matrix Mi,j, is $d_i = \sum_j M_{i,j}$. If we consider a subgraph V⊂G, to which node i belongs, we can split the total degree d in two quantities: di(V) = diin(V) + diout(V). The first term of the summation denotes the number of edges connecting node i to other nodes belonging to V (i.e., $d_i^{in}(V) = \sum_{j \in V} M_{i,j}$). The second term of the summation formula denotes the number of connections toward nodes in the rest of the graph (i.e., $d_i^{out}(V) = \sum_{j \notin V} M_{i,j}$). The first definition of communities is due to Flake (Flake, Lawrence & Giles, 2002a; Flake et al., 2002b), who defined a community as the set of nodes C (C⊂G) such that $d_i^{in}(C) \geq d_i^{out}(C)$. This definition was later improved to handle some degenerate cases in Ino, Kudo, and Nakamura (2005). Sidiropoulos et al. (2007) provided a looser definition of communities requiring only that the sum of all degrees of nodes within a community C is larger than the sum of all degrees toward the rest of the graph G (i.e., $\sum_{i \in C} d_i^{in}(C) > \sum_{i \in C} d_i^{out}(C)$). In general, we may give many different quantitative definitions of a community, which depend on the context of the application where it is developed.

Figure 2. Sample Web graph



The structure of a community can be encountered at various scales in the Web. The most thoroughly investigated are the inter-site communities, which span several Web sites, and usually define broad thematic areas determined by a set of keywords (e.g., the 9/11 community) (Flake, Tarjan, & Tsioutsiouliklis, 2004). The notions of compound documents (Dmitriev, Lagoze, & Suchkov, 2005; Eiron & McCurley, 2003) and logical information units (Li, Candan, Vu, & Agrawal, 2002; Tajima, Hatano, Matsukura, Sano, & Tanaka, 1999) are closely related to the Web communities, but at a much smaller scale, being comprised by a handful of Web objects of a single site, thus they are intra-site communities. A compound document is a logical document authored by (usually) one author presenting an extremely coherent body of material on a single topic, which is split across multiple nodes (URLs). Similarly, a logical information unit is not a single Web page, but it is a connected subgraph corresponding to one logical document, organized into a set of pages connected via links provided by the page author as "standard navigation routes." Sidiropoulos et al. (2007) extended the notion of intra-site communities and proposed communities whose topic is much more generic than the logical document's topic and their existence is determined by the density of the linkage among the pages they are comprised of. To support their claim, they examined several Web sites with a crawl available on the Web. As an intuitive step, they confirmed the existence of such communities using graph visualization. The visualization of all these networks was performed with the visualization package Pajek1. As a sample, they present the drawing of the http://www.hollins.edu Web site, whose January 2004 webbot crawl was available at the Web. The resulting image is illustrated at Figure 3. We can easily see the co-existence of compound documents (at the lower right corner), with compact node clusters (at the upper center), and less apparent clusters (at the upper right of the image).