

Clustering the Chilean Web

Satu Virtanen

Helsinki University of Technology
Laboratory for Theoretical Computer Science
P.O. Box 5400, 02015 HUT, Finland
Satu.Virtanen@hut.fi

Abstract

We perform a clustering of the Chilean Web Graph using a local fitness measure, optimized by simulated annealing, and compare the obtained cluster distribution to that of two models of the Web Graph. Information on web clusters can be employed both to validate generation models and to study the properties of the graph. Clusters can also be used in semantics-based grouping of websites or pages e.g. for indexing and browsing.

1. Introduction

In this work we examine the Chilean Web Graph (CWG), a coherent subset of the practically unobtainable Web Graph [4]. Many properties of the Web Graph have already been studied in detail (cf. [4]), but the cluster formation process underlying the network construction is relatively unknown. Identification of clusters as well as understanding the cluster structure of the Web helps to improve indexing, for example. It also allows for comparison of generation models proposed to match the Web Graph with respect to measures other than the more standard measures of e.g. path length and degree distributions that are often included in the models by design.

The CWG is a directed graph with websites as vertices; an edge $\langle u, v \rangle$ signifies that there exists a link from at least one webpage under u to a webpage under v . Links pointing to non-Chilean websites, pages from non-Chilean websites, self-loops, and edge multiplicities (that could be used as weights in clustering) were ignored in this work. The graph construction is based on a complete crawl of all Chilean websites in 2002 (cf. <http://www.todo.cl>), previously studied in [2]. Some measures of the crawl and the resulting graph are listed in Table 1.

Most of the components of the CWG are tiny; in addition to 6,904 single-vertex components, there is only one 2-vertex component in addition to the giant component \mathcal{G}

Number of Chilean websites	39,054
Number of webpages crawled	2,262,418
Number of webpages parsed	2,079,063
Number of edges in the CWG	1,357,662
Number of isolated vertices	6,904
Size of the giant component \mathcal{G}	32,148
Number of edges in \mathcal{G}	1,357,661

Table 1. Properties of the Chilean Web Graph

of order 32,148. Hence \mathcal{G} , in which each vertex is connected to at least one other included vertex, clearly dominates the CWG. In this work we report a clustering analysis conducted on \mathcal{G} .

Clustering in general is the process of organizing data into meaningful groups (see e.g. [8]). Each data element may either be considered to belong to one cluster or a hierarchy of clusters, depending on both the semantics of the data set and the clustering method employed.

In a web graph, such as the CWG, an intuitive cluster is a set of websites (or webpages, depending on the granularity of the graph construction) that are densely connected by hyperlinks, but have relatively few links to other parts of the web. Inherently global approaches for locating clusters [8] do not scale well for large graphs, as manipulating a large adjacency matrix is computationally demanding. Our clustering method is based on *local information* derived from the members of the clusters. We define a fitness value for each cluster candidate without referring to properties of websites not included in the cluster. This allows for optimization of the fitness measure in a local manner.

The local search heuristic employed in the reported clusterings is *simulated annealing* [1, 10]. Each cluster is determined by a stochastic examination of nearby vertices in the graph, likely to result in selecting a locally optimal cluster with respect to the fitness measure used for the simulated annealing. The global clustering obtained here for the

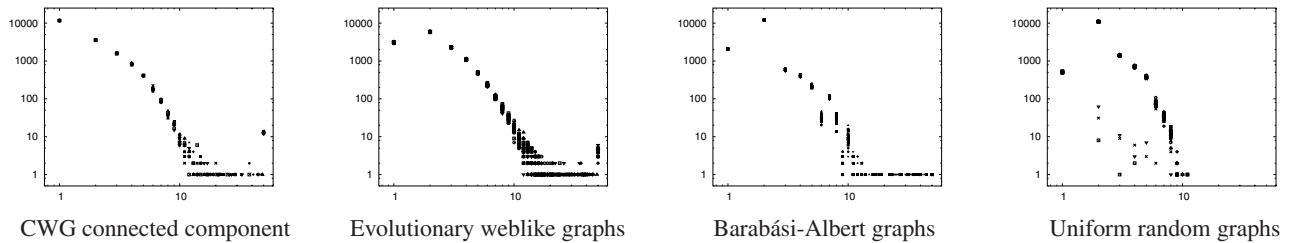


Figure 1. The cluster distributions (size vs. frequency) for \mathcal{G} (10 independent clusterings) and the comparison graphs (6 samples per model, each clustered 6 times). Cluster order was limited to 50; hence the cutoffs in the distributions.

CWG is flat and partitional, i.e., each vertex is assigned to exactly one cluster. This is achieved by excluding each locally optimal cluster already found from further clustering. The proposed method however also allows construction of hierarchical clusterings if instead of exclusion, the already located clusters are contracted to single vertices.

2. The fitness measure

A *cluster* \mathcal{C} in a directed graph $G = (V, E)$ is a set of vertices $\mathcal{C} \subseteq V$ that together induce a connected subgraph of G . The *internal degree* of a cluster is the number of edges that have both endpoints in \mathcal{C} : $\text{deg}_{\text{int}} = |\{\langle u, v \rangle \mid u, v \in \mathcal{C}, \langle u, v \rangle \in E\}|$. The *outward degree* of a cluster is the number of edges that have only the start vertex in \mathcal{C} : $\text{deg}_{\text{out}} = |\{\langle u, v \rangle \mid u \in \mathcal{C}, v \notin \mathcal{C}, \langle u, v \rangle \in E\}|$.

The *density* δ of a directed graph $G = (V, E)$ with $n = |V|$ vertices and $m = |E|$ edges is the ratio of m to the maximum number of edges possible, $n(n-1)$. Similarly, the *local density* δ_ℓ of a cluster \mathcal{C} , $\kappa = |\mathcal{C}|$, is the ratio of the internal degree of the cluster to the maximum possible:

$$\delta_\ell = \frac{\text{deg}_{\text{int}}}{\kappa(\kappa-1)}.$$

By convention, the density of an empty or a single vertex cluster is zero. Another usable fitness measure for clusters of *undirected* graphs is the *relative density*: the ratio of internal edges to the total number of edges incident on the cluster (see [13] and the references therein). For a directed graph, only edges with start vertices included in the cluster are locally available. Hence we redefine relative density δ_r for directed graphs by ignoring the edges that only have the end vertex in the cluster:

$$\delta_r = \frac{\text{deg}_{\text{int}}}{\text{deg}_{\text{int}} + \text{deg}_{\text{out}}}.$$

We want our clusters to have both high local and relative density, hence preferring “dense and introvert” clus-

ters as recommended in [9]. We use $f = \delta_\ell \cdot \delta_r$ as a cluster fitness measure, as it obtains high values when both δ_ℓ and δ_r are high, and low values if either one is low. A locally optimal cluster is one for which f cannot be increased by the removal of any included vertex or the addition of any neighboring vertex. A globally optimal cluster is a cluster with maximum fitness. A complete clustering of a graph can be obtained by locally optimizing f and iteratively either excluding the best cluster found from further clustering or contracting it to a single vertex, until all vertices are assigned to a cluster.

With the proposed method, also clusterings for partially unknown graphs are fluently obtained. In fact the method’s strongest asset is the possibility for on-line computation of clusters that are with high probability locally optimal with respect to the fitness measure. It is also noteworthy that the method requires no parameters such as a density thresholds or the requisite number of clusters. A more detailed discussion of the model is given in [14].

3. Clusters of the Chilean Web

We employed the method of the previous section to cluster \mathcal{G} , the connected component of the CWG. The cluster order was limited from above to 50 for computational ease. The initial clusters used to start the stochastic search contained up to 10 random neighbors of a random start vertex. At each step, one vertex was removed from or added to the cluster candidate, maintaining connectivity. The algorithm was repeated 30 times for each initial cluster, taking 250 modification steps per each iteration. From the clusters examined during this stochastic search, the one with the highest fitness was chosen as a cluster and excluded from further clustering.

The distribution obtained for \mathcal{G} is quite stable; the general shape and position of the distribution do not change over independent runs when varying the repetition counts and the order cutoff. We also examined graphs of the same

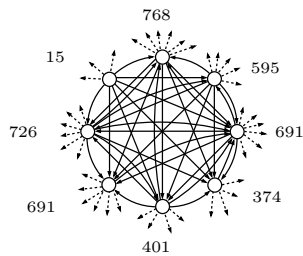


Figure 2. An example cluster found in one of the ten stochastic clusterings of \mathcal{G} with the number of outside links shown. Five of the websites are *.gotolatin.com domains, which explains their interconnectivity.

order and similar density as \mathcal{G} generated by two web-like graph models: the simple evolutionary weblike generation model of [12] (using uniformly distributed sources of the incoming links), and the Barabási-Albert construction for scale-free networks [3] (using a small uniform random graph as the seed graph and total degrees in preferential attachment). For comparison, we also studied the uniform random graph model of Erdős and Rényi [5] with the same order and size as the CWG. The resulting distributions are shown in in Figure 1.

As the CWG is a subset of the complete Web Graph, it is likely that some of the properties of the CWG can be generalized to the Web Graph. As the cluster distribution of \mathcal{G} may be characteristic of the Web Graph as well, we may assess the validity of weblike graph generators by the obtained distributions; an assessment of Internet graph generators by clustering is reported in [13]. Also the number of bipartite cliques in graphs has been used for theoretical model assessment in [11] for weblike graphs. According to Figure 1, the evolutionary model provides the best match to \mathcal{G} .

Also semantic conclusions of web structure can be drawn from a clustering of a subgraph of interest: websites on related topics are intuitively more likely to point to each other than sites covering other areas. This allows identification of topics and communities on the web [6, 7].

Acknowledgments

We thank Barbara Poblete from the University of Chile for providing the annotated and preprocessed crawl data and the anonymous reviewers for their comments.

The author is supported by the Academy of Finland under grant 81120 and the Helsinki Graduate School in Computer Science and Engineering.

References

- [1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Optimization and Neural Computing*. John Wiley & Sons, Inc., Chichester, UK, 1989.
- [2] R. Baeza-Yates, B. J. Poblete, and F. Saint-Jean. Evolución de la web chilena 2001–2002. Technical report, Centro de Investigación de la Web, Depto. de Ciencias de la Computación, Universidad de Chile, Santiago, Chile, Jan. 2003.
- [3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, Oct. 1999.
- [4] A. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1–6):309–320, June 2000.
- [5] P. Erdős and A. Rényi. On random graphs i. In *Selected papers of Alfréd Rényi*, volume 2, pages 308–315. Akadémiai Kiadó, Budapest, Hungary, 1976. First publication in Publ. Math. Debrecen 1959.
- [6] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, New York, NY, USA, 2000. ACM Press.
- [7] E. J. Glover, K. Tsioutsoulouklis, S. Lawrence, D. M. Pennock, and G. W. Flake. Using web structure for classifying and describing web pages. In *Proceedings of the 11th International World Wide Web Conference*, pages 562–569, New York, NY, USA, 2002. ACM Press.
- [8] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264 – 323, Sept. 1999.
- [9] R. Kannan, S. Vempala, and A. Vetta. On clusterings — good, bad and spectral. In *Proceedings of the 41st Annual Symposium on the Foundation of Computer Science*, pages 367–377, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.
- [10] S. Kirkpatrick, C. D. G. Jr. and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.
- [11] S. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the Web graph. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pages 57–65, Los Alamitos, CA, USA, 2000. IEEE Computer Society Press.
- [12] M. Levene, T. Fenner, G. Loizou, and R. Wheeldon. A stochastic model for the evolution of the web. *Computer Networks*, 39:277–287, 2002.
- [13] M. Mihail, C. Gkantsidis, A. Saberi, and E. Zegura. On the semantics of Internet topologies. Technical Report GIT-CC-02-07, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA, 2002.
- [14] S. Virtanen. Properties of nonuniform random graph models. Research Report A77, Helsinki University of Technology, Laboratory for Theoretical Computer Science, Espoo, Finland, May 2003.