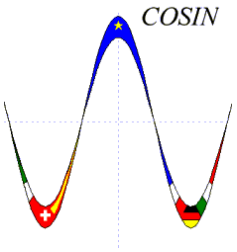

COSIN
IST–2001–33555

COevolution and **S**elf-organization
In dynamical **N**etworks

Algorithms for representing network centrality, groups and density and clustered graph representation

Deliverable Number:	D06
Delivery Date:	March 2004
Classification:	public
Contact Authors:	Marco Gaertler and Dorothea Wagner
Document Version:	final
Contract Start Date:	01-03-2002
Duration:	36 months
Project Coordinator:	Guido Caldarelli
Partners:	INFM *Italy* Università "La Sapienza" *Italy* Universitat de Barcelona *Spain* Université de Lausanne *Switzerland* Ecole Normal Superieure de Paris *France* Universität Karlsruhe *Germany*

Projects founded by the
European Commission under the
Information Society Technologies
Programme (1998-2002)

Abstract

One purpose of network analysis especially of social networks is to identify important actors, crucial links, subgroups, roles, network characteristics, and so on, to answer substantive questions about structures. There are three main levels of interest: the element, group, and network level. On the element level, one is interested in properties (both absolute and relative) of single actors, links, or incidences. Examples for this type of analyses are bottleneck identification and structural ranking of network items. On the group level, one is interested in classifying the elements of a network and properties of subnetworks. Examples are actor equivalence classes and cluster identification. Finally, on the network level, one is interested in properties of the overall network such as connectivity or balance. Algorithmic aspects concern the efficient computation of centrality indices, of groups and clusters, density, Clustering coefficient and transitivity. For this deliverable, algorithms for indices forming the basis of most studies were developed and implemented. A focus is on more efficient algorithms and robust and flexible implementations.

In many experimental studies, network indices computed for real world data are compared with randomly generated graphs satisfying certain properties. As the networks under consideration are large, efficiency is again a crucial issue. Therefore, new efficient generators have been designed and implemented to create graphs according to popular stochastic models such as random graphs, small worlds, and evolving graphs with preferential attachment.

Element Level – Vertex Indices

The starting point for this deliverable is the efficient computation of network indices on the element level. Algorithms for vertex indices forming the basis of most studies are developed and implemented. Most implementations are integrated in *visone*, a tool that facilitates the visual exploration of social networks [8]. The complete list of currently implemented indices is given in Figure 1. Moreover, a robust and flexible library of algorithms for vertex indices is developed based on the algorithms and data structure library LEDA [19]. This tool is described in [15]. It consists in suitable design patterns and basic algorithms for the most popular network indices, as e.g. the centrality and status indices listed in Figure 1. With methods from discrete statistics, we are able to visualize and summarize these results for large networks. Furthermore, we talk about heuristical speed-up and approximation techniques to handle even networks that are too large to get the correct results in an acceptable time.

A goal for future work in this deliverable is the design of unified and more efficient algorithms for network indices. For betweenness centrality, in particular, a substantial improvement over previous algorithms has already been achived. In [6] a new and more efficient algorithm is presented that computes betweenness centrality in time $O(nm)$ (instead of $O(n^3)$) by solving an augmented single-source shortest path problem from each vertex. Note that this is a significant speed-up for sparse graphs and thus for many real world data.

Group Level – Clustering

Clustering is a common technique to analyze and explore large data sets. Its main purpose is the identification of natural groups within the data. In the special case of graph clustering and network clustering, respectively, these groups represent grouped entities. For example, single web sites are merged into sets of common topics, or individual actors are abstracted to communities having the same interests and behavioral pattern. Although the notion of clustering seems to be clear with respect to our intuition, various different formal concepts have been proposed. The understanding of quality measurements and comparison techniques is still at a very basic level. Moreover, there exists no conclusive evaluation of algorithms that focuses on these aspects. For this deliverable we studied fundamental problems as well as real-world applications.

Experimental Evaluations

As a first step towards a better understanding of clustering and quality concepts, we concentrated on indicators based on intra-cluster density and inter-cluster sparsity. We presented the most important indices and an experimental evaluation in [7]. This benchmark suit tests already-known algorithms as well as our own one. The results showed clearly that every index had some weaknesses and could not be used for quality measurements on its own. However, we observe that the combination of several indices could cope them quite

index	definition	reference
local measures		
degree	$c_v = \sum_{e \in \text{instar}(v) \cup \text{outstar}(v)} \omega(e)$	–
indegree	$c_v = \sum_{e \in \text{instar}(v)} \omega(e)$	–
outdegree	$c_v = \sum_{e \in \text{outstar}(v)} \omega(e)$	–
distance measures		
betweenness	$c_v = \sum_{s \neq v \neq t \in V} \frac{\sigma_G(s,t\|v)}{\sigma_G(s,t)}$ where $\sigma_G(s,t)$ and $\sigma_G(s,t\|v)$ are the number of all shortest st-paths and those passing through v	[1, 11, 6]
closeness	$c_v = \frac{1}{\sum_{t \in V} \delta(v,t)}$	[4, 22]
eccentricity	$c_v = \frac{1}{\max_{t \in V} \delta(v,t)}$	[16]
radiality	$c_v = \frac{\sum_{t \in V} (\text{diam}(G) + 1 - \delta(v,t))}{(n-1) \cdot \text{diam}(G)}$	[23]
feedback measures		
status	$c_v = \alpha \cdot \sum_{(u,v) \in \text{instar}(v)} (1 + c_u)$ where $\alpha = \min\{\max_{v \in V} \text{indeg}(v), \max_{v \in V} \text{outdeg}(v)\}^{-1}$	[17]
eigenvector	$c_v = \mu^{-1} \sum_{(u,v) \in \text{instar}(v)} \omega(u,v) \cdot c_u$ where μ is the largest eigenvalue of $A(G)$	[5]
pagerank	$c_v = \gamma \cdot \frac{1}{n} + (1-\gamma) \sum_{(u,v) \in \text{instar}(v)} c_u$ where $0 < \gamma < 1$ is a free parameter	[9]
authority	$c_v = \mu^{-1} \cdot \sum_{(u,v) \in \text{instar}(v)} \omega(u,v) \cdot \sum_{(u,w) \in \text{outstar}(u)} \omega(u,w) c_w$ where μ is the largest eigenvalue of $A(G)^T A(G)$	[18]
hub	$c_v = \mu^{-1} \cdot \sum_{(v,w) \in \text{outstar}(v)} \omega(v,w) \cdot \sum_{(u,w) \in \text{instar}(w)} \omega(u,w) c_u$ where μ is the largest eigenvalue of $A(G) A(G)^T$	[18]

Figure 1: Available vertex centralities. Note that most indices have been generalized with respect to the original references, and all are rescaled to percentages.

well. Another interesting fact is that different algorithms behaved differently with respect to the density of the graphs. It seems that some algorithms are more suitable for sparse graphs than dense graphs and vice versa. Although graphs retrieved from real-world problems are typically sparse, complex networks often contain very dense parts as well. Thus algorithms are needed that can handle a large variety of densities.

Topics in News

It is usually very hard to find data sets that possess a already-known ideal clustering. In [10] we studied networks that consist of articles form the Wall Street Journal together with a similarity relation. The news are partial classify (by humans) with respect to different topics. Thus giving us the opportunity to compare our clusterings with an ideal one. Our algorithm (introduced in [7]) is very compatible with the given classification, i.e. many topics are identically found, some topics are merged and others split into many. Altogether, the results seem to be promising and can lead to an automatic classification scheme for news articles.

Clustering the graph of Autonomous Systems

In [12] we applied clustering techniques from [7] to the graph of the Autonomous Systems. Like other researchers before we found groups that reflect geographic and business-interests issues, see [14] for an example. However we focused more on dynamic aspects and associated effects. See also deliverable D17 for more details on this topic.

Graph Level – Computing Clustering-Coefficient and Transitivity

Since its introduction in [24], the clustering-coefficient has become a frequently used tool for analyzing graphs. In [20] the transitivity was proposed as an alternative to the clustering-coefficient. However, as we illustrate in [21] by several examples both parameters may differ vastly. On the other hand, an extension of the definitions to weighted versions provides the formal relation between them. It should be also mentioned that there is some variation in the literature with respect to nodes of degree less than two. Sometimes the clustering-coefficient is defined to be either zero or one. Alternatively, those nodes are not taken into account. However, the choice of the definition is important as can be seen from our results for the AS-graph. As many networks considered in complex systems are huge, the efficient computation of such network parameters is crucial. Several algorithms with polynomial running time can be derived from results known in graph theory. The main contribution of our work [21] is a new fast approximation algorithm for the weighted clustering-coefficient which also gives very efficient approximation algorithms for the clustering-coefficient and the transitivity. By an experimental study we demonstrate the performance of the proposed algorithms on real-world data as well as on generated graphs. These results also support the assumption that normally the values of clustering-coefficient and the transitivity differ considerably.

Generators

Random networks are frequently generated, for example, to investigate the effects of model parameters on network properties or to test the performance of algorithms. Recent interest in statistics of large-scale networks results in a growing demand for network generators that can generate large numbers of large networks quickly. Therefore, new efficient generators have been designed and implemented to create graphs according to popular stochastic models such as random graphs [13], small worlds [24], and evolving graphs with preferential attachment [2]. Time and space complexity of these generators is only linear in the size of the graph generated, and they are easily implemented [3].

References

[1] J. M. Anthonisse. The rush in a directed graph. Technical Report BN 9/71, Stichting Mathematisch Centrum, Amsterdam, October 1971.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] V. Batagelj and U. Brandes. Efficient generation of random graphs. To be presented at SUNBELT, 2004.

[4] M. A. Beauchamp. An improved index of centrality. *Behavioral Science*, 10:161–163, 1965.

[5] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120, 1972.

[6] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[7] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In *Proceedings of the 11th Annual European Symposium on Algorithms (ESA'03)*, LNCS 2832, pp. 568–579, Springer-Verlag, 2003.

[8] U. Brandes and D. Wagner. visone Analysis and visualization of social networks. In: P. Mutzel and M. Jünger (eds.) *Special issue on Graph Drawing Software*, Mathematics and Visualization, Springer, 2003.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[10] Steven R. Corman, M. Gaertler, and D. Wagner. Geometric MST clustering of text network collections. to be presented at SUNBELT 2004.

[11] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

[12] M. Gaertler and M. Patrignani. Dynamic analysis of the autonomous system graph. To appear in *Proceedings of IPS 2004, International Workshop on Inter-domain Performance and Simulation*.

[13] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[14] C. Gkantsidi, M. Mihail, and E. Zegura. Spectral analysis of internet topologies. In *IEEE Infocom 2003*, 2003.

[15] C. Gulden. Algorithmic Analysis of Large Network by Computing Structural Indices University of Konstanz, Master Thesis, 2004

[16] P. Hage and F. Harary. Eccentricity and centrality in networks. *Social Networks*, 17:57–63, 1995.

[17] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[18] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, 46(5):604–632, 1999.

[19] K. Mehlhorn and S. Näher. *The LEDA Platform of Combinatorial and Geometric Computing*. Cambridge University Press, 1999.

[20] M. E. J. Newman, S. H., Strogatz and D. J. Watts. Random graph models of social networks. *Proceedings of the National Academy of Science of the United States of America, vol. 99* pp. 2566–2572, 2002.

[21] T. Schank and D. Wagner. Computing Clustering-Coefficient and Transitivity, Faculty of Informatics, University Karlsruhe, Preprint 2004

[22] G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[23] T. W. Valente and R. K. Foreman. Integration and radiality: Measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105, 1998.

[24] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998.