Reconstruction of Networks from Their Betweenness Centrality*

Francesc Comellas and Juan Paz-Sánchez

Departament de Matemàtica Aplicada IV, Universitat Politècnica de Catalunya
Avda. Canal Olímpic s/n, 08860 Castelldefels, Catalonia, Spain
{comellas,juan}@ma4.upc.edu

Abstract. In this paper we study the reconstruction of a network topology from the values of its betweenness centrality, a measure of the influence of each of its nodes in the dissemination of information over the network. We consider a simple metaheuristic, simulated annealing, as the combinatorial optimization method to generate the network from the values of the betweenness centrality. We compare the performance of this technique when reconstructing different categories of networks –random, regular, small-world, scale-free and clustered–. We show that the method allows an exact reconstruction of small networks and leads to good topological approximations in the case of networks with larger orders. The method can be used to generate a quasi-optimal topology for a communication network from a list with the values of the maximum allowable traffic for each node.

1 Introduction

In recent years there has been a growing interest in the study of complex networks, related to transportation and communication systems (WWW, Internet, power grid, etc.), see [1,2]. Many of these networks are large with a number of nodes very often in the thousands. To store the topological details of the network requires knowing the list of adjacencies and, although usually the networks are sparse, this means the use of a large amount of memory. In contrast, many invariants of the network (degree sequence, eccentricity, spectrum, betweenness, etc.) contain important information with significantly less memory use. Therefore it would be of interest to reconstruct, even partially, a network from one (or more) of these invariants. Another related problem is the construction of a new network from a list of desired values of some relevant parameter associated to its nodes. One useful case would be the generation a topology for a quasi-optimal communication network from the values of the maximum allowable traffic for each node. In [3], Ipsen and Mikhailov use simulated annealing with an elaborated cost function based on the spectral density to perform such

* Research supported by the Ministerio de Educación y Ciencia, Spain, and the European Regional Development Fund under project TEC2005-03575 and by the Catalan Research Council under project 2005SGR00256.

M. Giacobini et al. (Eds.): EvoWorkshops 2008, LNCS 4974, pp. 31–37, 2008.

a reconstruction from the values of the Laplacian spectrum. Here we propose a reconstruction of a network topology from the values of its (vertex) betweenness centrality, a measure of the influence of each of its nodes in the dissemination of information over the network. The use of a simple cost function, together with the information provided implicitly by the knowledge of the betweenness centrality, drives the simulated annealing optimization method towards a good network reconstruction. The method is probabilistic, i.e. it contain a random component, and as a consequence we can not guarantee that the algorithm will find an optimal reconstruction, but we show that the final networks match the originals in their main topological properties.

In the next section, we introduce the mathematical notation and concepts necessary for this study, including a short description of simulated annealing, the combinatorial optimization technique considered here. Our main results are presented in Section 3.

2 The Betweenness Centrality of a Network and Its Reconstruction

We model a network as a graph $G = G(V, E)$, with vertex set V (order $n = |V|$) and edge set E.

Vertex betweenness or betweenness centrality (BC) was first proposed by Freeman [4] in 1977 in the context of social networks and has been considered more recently as an important parameter in the study of networks associated to complex systems [5,2]. BC is usually defined as the fraction of shortest paths between all vertex pairs that go through a given vertex. To be more precise, if $\sigma_{uv}(w)$ denotes the number of shortest paths (geodetic paths) from vertex u to vertex v that go through w, and σ_{uv} is the total number of geodetic paths from u to v, then we define $b_w(u, v) = \sigma_{uv}(w)/\sigma_{uv}$ and the betweenness centrality of vertex w is $b_w = \sum_{u,v \neq w} b_w(u, v)$. The normalized betweenness centrality of vertex w is defined as $\beta_w = \frac{1}{(n-1)(n-2)} \sum_{u,v \neq w} b_w(u, v)$, see Fig. 1.

In this paper, when we refer to betweenness centrality or BC we mean the set of values $\{\beta_1, \beta_2, \ldots, \beta_n\}$. The average normalized betweenness of a graph of order n is $\overline{\beta} = (\sum_{u \in V} \beta_u)/n$ and it is related to its average distance $\overline{l}$ as $\overline{l} = (n - 2)\overline{\beta} + 1$, see [6].

Here, we study the reconstruction of graphs from their BC. Note that the number of different graphs of a given order n is large even for relatively small

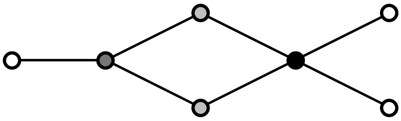

Fig. 1. The vertices of this graph have been colored according to their normalized betweenness centrality value: white 0, light grey 1/5, dark grey 11/30, black 19/30

orders. For example, for $n = 40$ there are approximately 10^{186} graphs. It makes no sense to check all of these graphs to find one with a matching BC, even in an approximate way. We are in the classical situation where combinatorial optimization algorithms (simulated annealing, genetic algorithms, tabu search, ant colony based systems, etc.) are useful, see [7].

For this initial study, we have considered as optimization method a standard version of simulated annealing (SA) [8]. As it is known, this method is inspired in the analogy made between the states of a physical system, e.g. a liquid, and the configurations of a system in a combinatorial optimization problem. A controlled heating/cooling process of the liquid (annealing) results in a true crystal (a minimum energy state) and avoids reaching a disordered glassy state. In the analogy, a change that decreases the cost of a function, ϵ, which measures the quality of a graph topology (see below), is always accepted, whereas if the cost increases, the change is accepted with a certain probability $e^{-\Delta\epsilon/T}$. (T is a control parameter known as temperature because of the analogy.) At a given temperature, a number of attempts N, large enough to obtain a good statistical set of trials, is performed and thereafter the temperature decreased. This process is repeated and the system is gradually cooled until it is stopped according to some criteria (time, number of changes accepted, etc.) In pseudo-code the SA algorithm can be written as follows:

```
1. Generate an initial random graph. Fix the initial value of $T$ and $T_{min}$.
2. Repeat $N$ times.
   (a) Modify the graph topology and find new cost.
   (b) If better, accept it as current solution.
   (c) If worse, accept only if $e^{-\Delta\epsilon/T} > rand()$
3. Lower $T$ and repeat 2 until $T < T_{min}$ or other stop criterion.
```

In our case we will reconstruct a given reference graph G_0 from its BC, $\{\beta_1^0, \beta_2^0, \ldots, \beta_n^0\}$. To perform a reconstruction we generate an initial random connected graph with n vertices (each vertex $v \in V(G)$ has random degree δ_v, $1 \leq \delta_v \leq n-1$). During the simulated annealing process, a typical graph modification consists of reconnecting all the edges of one vertex chosen at random. This reconnection is performed by deleting all the edges of this vertex and introducing r, $1 \leq r \leq n-1$, new random edges avoiding duplicate connections and ensuring that the new modified graph is also connected. To decide if the changes should be accepted, we need a measure (cost function) of the "distance" of a given graph G_t with BC $\{\beta_1^t, \beta_2^t, \ldots, \beta_n^t\}$ to the reference graph G_0. We introduce a simple distance function based on the quadratic difference of the BC, $\epsilon = \sum_{i=1}^{n}(\beta_i^0 - \beta_i^t)^2$. This function, suggested by the least squares method, is a natural choice in some multivariable optimization problems. On the other hand, we have tested other related functions assigning weights to the BC elements, but they are more complex and their efficiency is similar.

The main problem with the reconstruction of a graph is to relate the final graph with the reference graph. The use of a vertex graph invariant in reconstructing a network might be hampered by the degeneracy of almost all known graph invariants, and in our case two or more topologically distinct vertices might have identical betweenness values. Moreover, the reconstructed graph can be isomorphic to the original graph but with permuted vertices or non-isomorphic with some topological similarity that might not be manifest. Although this is an important question when reconstructing a graph from its spectrum, in our case and as each value of the BC is directly associated to a vertex, the problem only appears when the BC contains several entries with the same value.

As in [3], we check graph similarity using the singular value decomposition of the adjacency matrices of the reference and final graphs . We recall that a matrix A can be decomposed into two matrices U and V and a diagonal singular value matrix Σ which satisfy $A = U\Sigma V^T$ and $\Sigma = U^T AV$. For any two graphs G_1 and G_2 with adjacency matrices A_1 and A_2, consider the function $F = F(A_1, A_2) = U_1\Sigma_2 V_1^T = U_1 U_2^T A_2 V_2 V_1^T$ which is constructed from the singular vectors of G_1 and G_2. If the two graphs are isomorphic and their adjacency matrices only differ because of a different ordering of the vertices, it will happen that $A_1 = F(A_1, A_2)$. However, if the two graphs are not isomorphic, F will have real values not far from the values of A_1. Therefore, it is possible to define $\Delta = A_1 - F$ and use the norm $\delta = \sqrt{\sum_{i,j} \Delta_{ij}^2 / n}$ to measure similarity between the graphs.

We note that two isomorphic graphs have the same BC, which is independent of the labeling of the vertices, but there also exist non-isomorphic graphs (topologically different) with the same BC, which we call *isobet* graphs. For $n \leq 5$ there are no connected isobet graphs. For $n = 6$ there exist two pairs, there are 15 pairs for $n = 7$, etc. The number of isobet graphs increases rapidly with the order of the graph, but the fraction is very small. Hence, two graphs with the same BC would indeed be isomorphic with a high probability.

To know if two graphs are isomorphic is a difficult problem. It has been proved to belong to the class NP but it is thought not to be an NP-complete problem [9]. There is no known efficient (polynomial time) algorithm to solve this problem. Schmidt and Druffel [10] propose the method which we have implemented in our study. Their algorithm is not guaranteed to run in polynomial time, but has been shown to perform efficiently for a large class of graphs. Two isomorphic graphs should have the same exact degree distribution. After checking this property, the Schmidt and Druffel algorithm uses information from the distance matrices of the graphs to establish an initial vertex partition. Then, the distance matrix information is applied in a backtracking procedure to reduce the search for possible mappings between the vertices of the two graphs. The algorithm returns this mapping if the original and reconstructed graphs are indeed isomorphic.

3 Results and Conclusion

The SA algorithm was implemented in C++ (Xcode) and executed on an Apple Xserver G5 with dual PowerPC processors at 2.3 GHz. The parameters considered for the SA are $T_0 = 1.0$, $N = 2000$, $T_{min} = 0.000001$ and a geometric cooling rate $T_{k+1} = 0.9T_k$.

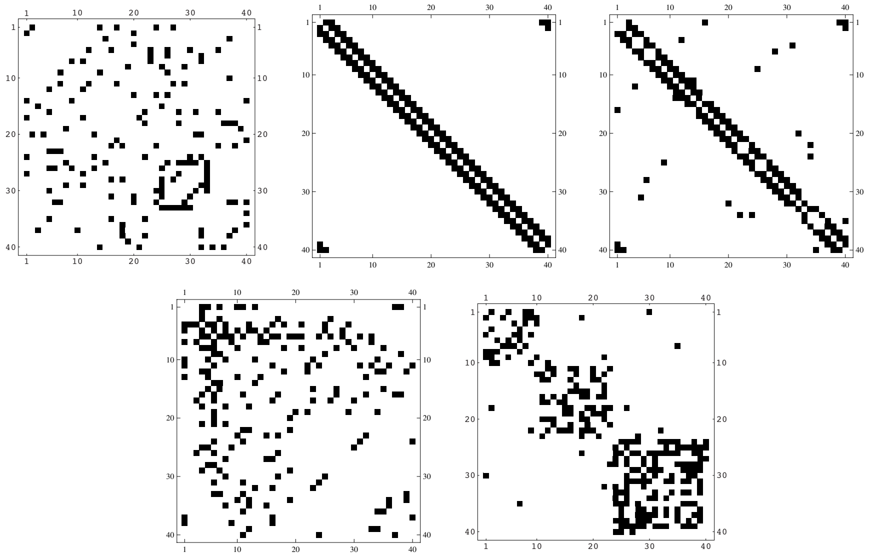

Fig. 2. Adjacency matrices of the reference graphs: random, regular (circulant), Watts-Strogatz small-world, scale-free and clustered. All graphs haver order 40.

The main study was performed as follows:

We generate one sample graph of order 40 for each of the categories considered: random, regular (circulant), Watts-Strogatz small-world [11], scale-free [1], and clustered. Fig. 2 shows a graphic representation of the adjacency matrices of these reference graphs. For each reference graph, we compute the betweenness centrality and use it to reconstruct the graph with the simulated annealing method. To be fair in the comparisons, we fix the reconstruction time for each graph to be 900 seconds. After this time, we compute the main topological parameters (diameter, average distance, degree distribution, clustering) for the best graph obtained and we check the similarity of its adjacency matrix with the original graph. Each test is repeated 500 times and the results are averaged. Fig. 3 shows a typical reconstruction.

In Table 1, we present a set of results for this method. We can see that the reconstruction gives acceptable results in all cases, but provides better approximations for graphs with some randomness in their structure, and such that their vertices have different betweenness centrality values. This is the case, obviously, of random graphs and also scale-free graphs.

We also tested the algorithm using graphs with small orders (up to 12 vertices) and in all cases we were able to obtain an exact reconstruction of the graph.

The results show that a simple metaheuristic method, simulated annealing, with an also simple cost function, reconstructs small graphs exactly from their betweenness centrality and obtains a topologically good approximation for larger

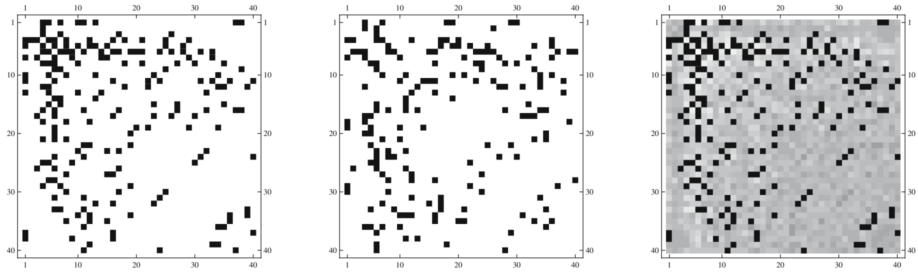

Fig. 3. Reconstruction of a scale-free graph using simulated annealing. Left: The adjacency matrix of the original graph. Center: The adjacency matrix of the reconstructed graph. Right: Matrix F of the reconstructed graph, see Section 2.

Table 1. Simulated annealing. Results for the average of 500 reconstructions for each reference graph. Graph order= 40, $T_0 = 1.0$, $N = 2000$, $T_{min} = 0.000001$, geometric cooling rate $T_{k+1} = 0.9T_k$.

		Random		Circulant		Small-world		Scale-free		Clustered	
		Ref.	*Recns.*	Ref.	*Recns.*	Ref.	*Recns.*	Ref.	*Recns.*	Ref.	*Recns.*
Diameter	avg.	6	5.870	10	8.770	6	6.687	4	4.475	5	5.271
Avg. Dist.	avg.	2.89	2.780	5.38	4.670	3.31	3.028	2.32	2.340	2.65	2.578
Degrees	min.	1	1.000	4	1.999	3	1.000	1	1.11	2	3.75
	avg.	3.8	3.950	4	2.570	4	3.690	4.95	5.096	6.3	4.464
	max.	8	10.000	4	3.870	5	6.190	17	15.760	13	14.650
Clustering	avg.	0.2	0.195	0.5	0.030	0.32	0.122	0.26	0.262	0.37	0.200
Norm. BC	avg.	0.050	0.047	0.115	0.097	0.061	0.053	0.035	0.035	0.043	0.042
δ	avg.		0.026		0.074		0.029		0.024		0.082

graphs. (We have tested graphs with up to 2000 nodes and 20000 [12].) The method works without modification in the related problem of the construction of a new network from a a list of desired values for the maximum allowable traffic for each node.

Our extensive tests show that the cost function considered, the quadratic difference of the BC, is a good choice for SA, and the method is a nice alternative to the reconstruction from the Laplacian spectrum as it is easier to implement and results in a faster algorithm, allowing reconstructions of similar quality.

Further work is planned to evaluate and compare the performance of other combinatorial optimization methods, like ant colony optimization [13], multiagent systems [14], tabu search [7,15] and genetic algorithms.

References

1. Barabasi, A.-L., Bonabeau, E.: Scale-free networks. Scientific American 288(5), 50–59 (2003)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
3. Ipsen, M., Mikhailov, A.S.: Evolutionary reconstruction of networks. Phys. Rev. E. 66, 46109 (2002)
4. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry 40, 35–41 (1977)
5. Goh, K.-I., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of scale-free networks. Proc. Natl. Acad. Sci. USA 99, 12583–12588 (2002)
6. Comellas, F., Gago, S.: Synchronizability of complex networks. J. Phys. A: Math. Theor. 40, 4483–4492 (2007)
7. Aarts, E., Lenstra, J.K. (eds.): Local Search in Combinatorial Optimization. John Wiley & Sons Ltd, New York (1997)
8. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
9. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, New York (1979)
10. Schmidt, D.C., Druffel, L.E.: A fast backtracking algorithm to test directed graphs for isomorphism using distance matrices. Journal of the ACM 23, 433–445 (1976)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. Nature 393, 440–442 (1998)
12. Paz-Sanchez, J.: Reconstrucció de grafs a partir del grau d’intermediació (betweenness) dels seus vèrtexs. PFC (Master Thesis) (in Catalan) (July 2007)
13. Dorigo, M., Stützle, T.: Ant colony optimization. MIT Press, Cambridge (2004)
14. Comellas, F., Sapena, E.: A multiagent algorithm for graph partitioning. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) EvoWorkshops 2006. LNCS, vol. 3907, pp. 279–285. Springer, Heidelberg (2006)
15. Glover, F.: Future paths for integer programming and links to artificial intelligence. Comput. & Ops. Res. 13, 533–549 (1986)