Propagation d-Covering on weighted complex networks

Alessio Cardillo
Department of Physics and Astronomy University of Catania,
Via S.Sofia 64, I-95123, Catania, Italy.
http://www.ct.infn.it/~cardillo/, alessio.cardillo@ct.infn.it

Abstract

The search for an efficient allocation of resources in complex networks is a topical issue concerning communication, transportation and social systems. Most of the proposed protocols are based on the degree distribution of the graph. However, the complex networks under study are better represented as weighted graphs (graphs in which relation between two nodes (a link) can assume any value). The Propagation d-Covering represent a heuristic method to find near-optimal solutions to the covering problem in real communication networks, taking into account links weights. In this paper the benefits of using HPC resources in order to obtain results are shown.

1 Introduction

In the recent years, big efforts have been spent on studying the effects of epidemic dynamics in real world systems [1, 2, 3]. The natural framework for studying such systems is graph theory and in particular *complex networks* theory. In fact, many systems surrounding us can be easly mapped into a graph. In this paper we will discuss about an application of immunization strategies over complex network. In particular we deal with the problem of finding a methodology to immunize a network while minimizing the amount of resources spent to do it. This because, in realistic cases we have to deal with the problem of immunization cost and knowledge of the system. The *Propagation d-Covering* is a heuristic method which tries to achieve good immunization taking into account the previous exposed issues. This methodology could be used to immunize, for example, computer network or social networks aganist different kinds of epidemic processes. In particular we discuss about the process of parallelization of the program in order to achieve better performance in terms of computation time.

2 The *Propagation d-Covering* algotithm

A highly topical problem is the development and deployment of an immune system to prevent technological (or not) networks from spreading viruses. In this case it is worthwhile to characterize whether a centralized organization or a distributed approach is the best. The solution to this and similar problems may be computationally easy or hard depending on the topological properties of the underlying graph. Following the idea of Echenique *et.al.*[5, 6, 7], using an heuristic algorithm that targets vertices, we compute an upper bound to the minimum fraction of nodes needed to cover a graph. This algorithm is called *Propagation d-Covering* and acts as follows:

1. We select at random a node i and we consider all its d-neighbours (neighbours at distance d) with $d = 1, 2, 3, \ldots, n$;
2. we select the one with the highest value of a certain quantity (i.e. degree, strenght, betweenness [8, 9, 10]) and *cover* it;
3. now we consider the d-neighbors of the *cover* node and consider them as *covered*;
4. repeat all the above operations for all the uncovered nodes until all the nodes in the network are *covered.*

All the operations described above can be seen in the following picture.

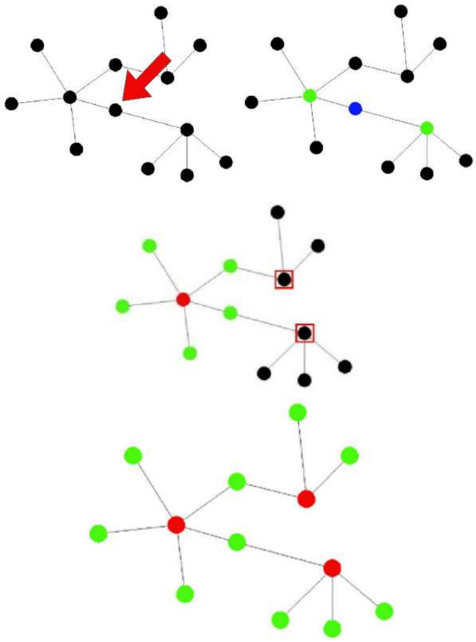

Figure 1: (color online) Operations made by the algorithm (from top left to bottom): Choice of the starting node; selection of its d-neighbours; Selection of: the cover node, of its d-neighbours and of nodes where to start the other iterations; final covering status of the net. All these pictures have covering distance d equal to 1.

2.1 Covering strategies

In order to decide if a node could be a cover or not several quantities can be taken into account. If we decide to ignore the "interaction" between nodes considering only the topology of the network, a good measure to perform covering is the *degree.*

$$k_i = \sum_{j=1}^{\mathcal{N}} a_{ij} \; ; \tag{1}$$

Where a_{ij} is the adjacency matrix. In this way we decide to immunize those nodes which are connected with the highest number of nodes possible (according to the algorithm procedures).

On the other hand, since almost all real networks are weighted, these weights must be taken into account when the covering algorithm is performed. This lead to the choice of *strenght* as parameter to be used. The strenght of node i is defined as:

$$s_i = \sum_{j=1}^{\mathcal{N}} a_{ij} \, w_{ij} \; ; \tag{2}$$

Where a_{ij} is the adjacency matrix and w_{ij} is the weight of the link $i - j$. Another possible quantity to look at is the *average link weight per node* w_i defined as:

$$\overline{w}_i = \frac{\sum_{j=1}^{\mathcal{N}} a_{ij} \, w_{ij}}{\sum_{j=1}^{\mathcal{N}} a_{ij}} \; ; \tag{3}$$

This is an important factor in order to understand how many packages passes through a node. In communication networks, this quantity could be a better estimator of node centrality than merely degree or strenght.

2.2 Dataset

The dataset used is a network obtained from the Gnutella P2P network [11]. The carachteristics of this network are reported in the table below.

Data	
$\mathcal{N}$	79939
$\mathcal{K}$	165059
k_{max}	15665
s_{max}	73412
$\overline{w}_{max}$	232.6

Table 1: Principal properties of the used dataset. Number of nodes $\mathcal{N}$, number of links $\mathcal{K}$, maximum degree k_{max}, strenght s_{max} and maximum average link weight per node $\overline{w}_{max}$.

3 High performance computing and parallelization

Since the the algorithm depends on the starting node, each simulation perform different covering realizations with different starting nodes. So the code can be easly parallelized dividing the total number of single covering realization upon different computing nodes (from two up to the number of realizations). In the context of parallel computing this kind of problems are called *intrinsically parallel* because they execute many times the same code and each sub-simulation is indipendent from the others. In this sense the best way to perform parallelization is through distributed memory parallelism and in particular using the Message Passing Interface (MPI) protocol. Theoretically splitting the simulation over n computing nodes allows to obtain a computing time which is equal to $\frac{1}{n}$ times the sequential one. However, following Amdahl's law the speed-up cannot exceed an upper bound. Thanks to MPI parallelization it has been possible to reduce the overall simulation time from about 15 hours to about 4 using just 4 processors. So the speed-up factor is equal to:

$$S = \frac{T_{old}}{T_{new}} = \frac{905'\,25.769''}{238'\,18.732''} = 3.8\,;$$

4 Results

To measure how covering acts on these networks, we decided to compute for each value of distance d two quantities: the *mean number of nodes served by a cover* $\langle N_{nc}\rangle$ (Number of nodes per cover). This number gives the average number of nodes at distance d from a cover and it is defined as:

$$\langle N_{nc}\rangle = \frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{j=1}^{\mathcal{N}} a_{ij}\ ; \tag{4}$$

Where N_c is the number of covers, $\mathcal{N}$ is the number of nodes and a_{ij} is the adjiacency matrix element. The other quantity is the *mean number of covers covering a node* $\langle N_{cn}\rangle$ (Number of covers per node).

$$\langle N_{cn}\rangle = \frac{1}{\mathcal{N}}\sum_{j=1}^{\mathcal{N}} \delta_j\ ; \tag{5}$$

Where δ_j is equal to 1 if j is a cover and zero otherwise. It is very important to notice that the higher this number is the better immunizzazion is. The reason is that an higher number of covers per node means that each node has more than one cover at distance d. One could expect that this number goes down as d increases. Instead, as could be seen in fig. 2, results shows a peak for a specific value of d. However this happens only in the case of degree based covering (k-covering). This means that there is an *optimal* value of d which ensures the best coverage over the net. This differentiation is not visible on nets of [13] and further analysis are required in order to understand that. Nonetheless, as expected, the plots showing the fraction of nodes covered by a cover as a function of d shows a smooth transition from 0 to 1. This because while the number of covers at each distance d decreases, the number of nodes covered by a single node icreases reaching the values of $\mathcal{N}$ for $d \simeq D$ where D is the *diameter* of the net.

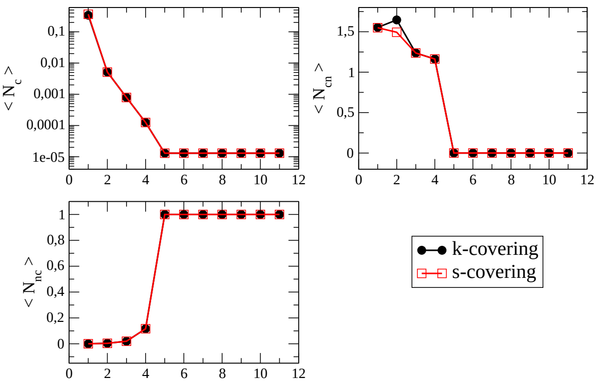

Figure 2: (color online) From left to right and top to bottom. Average number of covers $\langle N_c \rangle$, average number of covers per node $\langle N_{cn} \rangle$ and average number of nodes per cover $\langle N_{nc} \rangle$ as a function of d for both degree based covering (k-covering) and strenght based one (s-covering).

5 Conclusions

The possibility to use high-performance computing resources is a powerful tool for improving research quality. In particular, in the case of Propagation d-Covering the use of Marenostrum supercomputer lead to a drastic reduction of computing time. Saving time not only allowed to obtain the results quickly but also allowed to use the time conceeded in both starting further development of covering algorithm and also in the begining of collaboration on new topics. In particular for the former, since Propagation d-Covering results are not unique but depends on the starting node we are studying a methodology to indicate which covering realizations are better respect to the capability to slow down eventual epidemic process spreading through the net. For the latter during the visiting period togheter with Dr. Jesus Gòmez-Gardeñes a collaboration on application of evolutionary game theory on complex networks has started.

Aknowledgments

I want to thank professor Albert Díaz-Guilera for all the support given to me during my stay in Barcelona and for all the useful discussions and suggestions. I also want to thank my supervisor professor Vito Latora for all the support given and Dr. Jesus Gardeñes not only for his support but also for his friendship. Last, but not the least I want to thank all the staff of BSC for the kindness and the support given to me during my stay. This work was carried out under the HPC-EUROPA++ project (project number: RII3-CT-2003-506079) with the support of the European Commission Capacities Area - Research Infrastructures Initiative.

References

[1] V. Colizza, A.Barrat, M. Barthelemy, and A. Vespignani, Proceedings of the National Academy of Sciences USA, **103**, 2015-2020 (2006).

[2] M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Physical Review Letters **92**, 178701 (2004).

[3] V. Colizza, M. Barthelemy, A. Barrat, A.-J.Valleron, and A.Vespignani, PlOS-Medicine, **4**, e13 (2007).

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Physics Reports **424** (2006) 175

[5] P. Echenique , J. Gómez-Gardeñes , Y. Moreno and A. Vásquez , Phys. Rev. E **71** 035102(R) (2005).

[6] J. Gómez-Gardeñes , P. Echenique and Y. Moreno, Eur. Phys. J. B **49** 259-264 (2006).

[7] Y. Moreno , R. Pastor-Satorras and A. Vespignani, Eur. Phys. J. B **26** 521-529 (2002).

[8] U. Brandes, J. Math. Soc. **25** (2001) 163.

[9] K.I. Goh, B. Kahng, D. Kim, Phys. Rev. Lett. **87** (2001) 278701.

[10] S. Wasserman, K. Faust *Social Networks Analysis* (Cambridge University Press, Cambridge, 1994).

[11] F. Wang, Y.Moreno, and Y. Sun, Physical Review E , **73**, 036123 (2006).

[12] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein; *Introduction to Algorithms* 2-nd Edition, The MIT Press, Cambridge, US, (2001).

[13] A. Cardillo, J. Gómez-Gardeñes, V. Latora, Y. Moreno; *d-Covering on Weighted Complex Networks*, poster presented at the conference: *Complex Networks, from biology to information technology*, 2-8 July 2007, Pula (Italy).