SER-Based Web Graph Decontamination

Vanessa Carla F. Gonçalves[1], Felipe M. G. França[1], Nelson Maculan[1], Priscila M. V. Lima[2]

[1]Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

[2]Universidade Federal Rural do Rio de Janeiro, Seropédica, Brazil.

{vcarlaufrj,felipe}@cos.ufrj.br, priscilamvl@ufrrj.br

Abstract. *The creation of a web bubble, or link farm, is a technique used by web spammers to increase the visibility of a target page T. Specialized mobile agents, called Web Marshals, are employed to detect and dismantle link farms. As the number of Web Marshals, as well as the number of hops needed to disassemble a link farm, need to be minimized, this kind of process can be seen as a graph decontamination problem. This paper presents an asynchronous distributed decontamination algorithm which can be embedded into the web marshals' behavior. Although the novel algorithm is topology independent, compared to recent related works, having circulant graphs as targets, it presents equal or better performance and a smaller number of deployed web marshals when applied to the same circulant graph topologies.*

Keywords*: web graph, link farms, graph decontamination, scheduling by edge reversal, graph search, randomized distributed algorithms.*

1. Introduction

A *web bubble* or *link farm* is a set of web pages that point to a target page *T*. These farms are built by web spammers who want to increase the visibility of page *T*. Plenty of methods to hide links that make part of a farm have been designed in conjuction with mechanisms to avoid the rupture of these connections [Gyöngyi and Garcia-Molina 2005].

Web crawlers are employed in order to detect this kind of "infection" and several of them have being tested to increase the efficiency of the detection of such attacks. Many methods to dismantle link farms have being proposed in the literature [Luccio and Pagli 2007][Flocchini et. al. 2005][Flocchini et. al. 2007]. The "cure" for such infections have been known in the community as a graph search problem called *graph decontamination*. In that line, we can say that a web bubble or link farm is a contaminated graph. Pages belonging to a link farm are the nodes of that graph and the contamination is the presence of the link to the target page *T,* added by the spammers.

Although a contamination can be defined as a link to an elected page, a virus jumping between hosts, or an exploration team moving through a forest, here we restrict ourselves to the problem of link farms. In the latter case, decontamination consists in breaking the links that maintain the farm and make sure that a page will never become

contaminated again. In general, once a node is exposed to a contaminated neighbor, if it's not protected, it can be contaminated again. In the more strict case, the number of contaminated neighbors that can contaminate an unguarded node is equal to one (1). However, an acceptable generalization consists in consider that, in the case of link farms, only by obtaining a majority of contaminated neighbors a broken link can be restored [Luccio and Pagli 2007].

In order to find a way to perform this kind of "cleaning" in a set of web pages, special mobile agents, called *Web Marshals* (WMs) [Luccio and Pagli 2007] have been developed. Their job is to travel, node by node (or page by page), fixing the contamination they are designed to destroy. As circulant graphs are a typical structure used by web spammers, [Luccio and Pagli 2007] have proposed a distributed algorithm, working in synchronous and asynchronous modes, to be embedded into autonomous agents, i.e., WMs, in order to destroy link farms. Bounds for the number of WMs and for the number of hops were provided.

Our new distributed decontamination algorithm is based on the *Scheduling by Edge Reversal* (SER) graph dynamics [Barbosa 1996][Barbosa 2000], a resource sharing distributed algorithm. SER works as follows: starting from any acyclic orientation over the edges of any arbitrary target graph G, only sink nodes, i.e., nodes having all of their adjacent edges directed to themselves, are allowed to "operate" upon shared resources. After operating, sink nodes reverse the orientation of all adjacent (incident) edges, becoming source nodes. As new sink nodes are eventually defined, the dynamics is indefinitely preserved. SER had been applied to a large spectrum of applications, such as in the (i) design of asynchronous (clockless) digital circuits [Cassia et. al. 2009][França et. al. 2007]; (ii) integrated scheduling of Job Shop and AGVs (Automated Guided Vehicles) in flexible manufacturing systems [Lengerke et. al. 2008a][Lengerke et. al. 2008b]; (iii) design of collision free MAC protocols [Pinho et. al. 2009], and in (iv) biologically plausible rhythmic generators, such as CPGs (Central Pattern Generators) [Yang and França 2003][Braga et. al. 2008].

SER-based decontamination works in the following way: WMs are associated to sink nodes; once decontamination is performed at a sink node, new WMs are sent, via replication, only to immediate neighbor nodes that will become sinks upon termination, i.e., via edge reversal. The SER-based approach implicitly associate the amount of concurrency provided by a SER dynamics to the number of concurrently operating WMs and the total number of decontamination steps performed, i.e., one hop is counted each time a node receives one (or more) WM copy in order to become a new sink. Our approach produces the same or better quality solutions found by [Luccio and Pagli 2007] while capable of working in arbitrary connected topologies and under more strict contamination rules, what suggests its appropriateness on dealing with new kinds of web attacks.

This work is organized as follows. Section 2 presents related works. Section 3 explains the new distributed graph decontamination algorithm and Section 4 discusses our experimental results. Section 5 contains conclusion and future works.

2. Related Works

Luccio and Pagli [Luccio and Pagli 2007] present an algorithm to destroy a formed link farm with WMs, equivalent to the mobile agents used in [Flocchini et. al.

2005][Flocchini et. al. 2007] which are passed through the links that form the link farm and, once they get in a page, they detroy the link to the target page *T*.

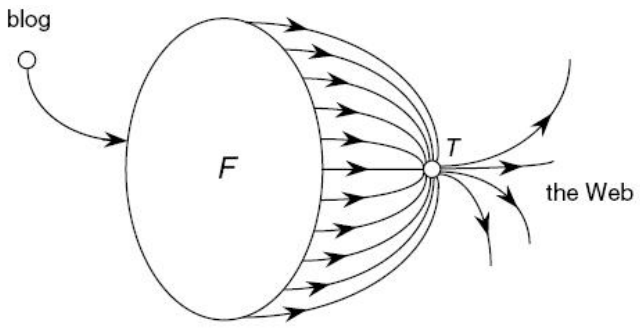
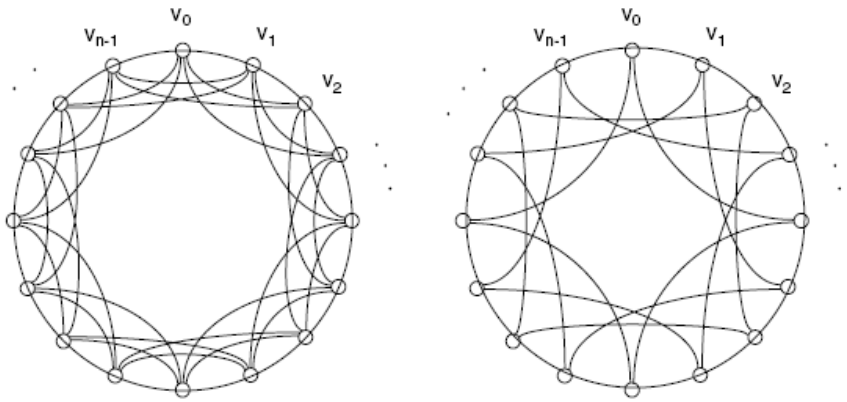


Figure 1. The structure of a link farm. F: Circulant Graph [Luccio and Pagli 2007].

It is assumed that a link farm *F* has a circulant graph as its main structure [Luccio and Pagli 2007] (see Figure 1) and their work is strongly based on the properties of this particular graph topology. It is shown that in a circulant graph $C_{i,n}(L)$, with *L* being a list of integers $\{1, 2..., k\}$ and *k* being the maximum integer in that list, $k + 2$ WMs are needed in order to extinguish the link farm. In Figure 2 (a) $k = 3$, and in Figure 2 (b) $k = 2$ ($k = |L|$, *L* being {1,4}).



(a) **(b)**

Figure 2. Examples of Circulant Graphs: (a) $C_{i,n}(1, 2, 3)$; (b) $C_{i,n}(1, 4)$ [Luccio and Pagli 2007].

A graph decontamination method based on the visibility of each node is proposed in [Flocchini et. al. 2007]. It is concluded that assuming visibility 2, i.e., when an agent can "see" the node hosting it, its immediate neighbors and the neighbors of its immediate neighbors, is needed to prevent the overuse of mobile agents. The overuse of mobile agents is one of the main concerns of the present work. Another concern of this work is about providing a topology independent distributed algorithm, i.e., a SER-Based decontamination able to deal with any given graph structure.

3. Edge Reversal Decontamination

From any acyclic orientation ω of the target graph *G*, it is possible to start a decontamination dynamics based on the SER behavior. It is easy to see that there

always exist a node coloring of G associated to ω in the following manner: each node receive color equal to the length of the longest directed path from it to a sink node; this means sink nodes in ω receive color $\lambda = 0$, and this is called the *sink decomposition* of ω. Starting by placing Web Marshals (WMs) into sink nodes (*home bases*), it is also easy to see that nodes having color $\lambda = 1$ are next to turn into color $\lambda = 0$ (sinks) upon edge reversal of sinks.

Since having sink decompositions of large length increase the probability of dealing with fewer concurrent sink nodes, i.e., minimize the number of WMs in the web graph, the design of a heuristics able to do it in web graphs is one of the goals of this work. Having in mind the particular properties of circulant graphs, this work also aims to compare the performance, both in terms of the number of WMs and the number of steps taken to decontaminate, between the algorithm proposed by and the SER-based strategy introduced here. The next two subsections presents (i) the *Alg-Stretcher*, an algorithm to enlarge the length of sink decompositions of already defined acyclic orientations produced by *Alg-Edges* [Arantes Jr et. al. 2009], a randomized distributed algorithm designed to produce acyclic orientations over anonymous networks, and; (ii) the *Alg-Decontamination* algorithm, the edge reversal based distributed algorithm; applied, in this work, to circulant graphs used to build link farms.

3.1 Alg-Stretcher

Let $\lambda = l_{max}$ be the outer layer in the sink decomposition of a target acyclic orientation ω. From layers $\lambda = (l_{ma} - 1)$ to $\lambda = 0$ in the sink decomposition, each node v is tested about being moved to a new outer layer $\lambda > l_{max}$. If an increase in the number of layers of the sink decomposition of the graph is obtained, the resulting acyclic orientation ω' is obtained by having all of v's edges oriented according to the direction of the sink decomposition of the graph, otherwise (no increase in the sink decomposition) the previous orientation is kept.

Experimental results from the application of *Alg-Stretcher* in the acyclic orientations produced by *Alg-Edges* are depicted in Figure 3. Each point represents the mean value of 500 runs over connected random graphs. Apart from being the algorithm producing the largest number of colors, compared to other two distributed heuristics [Arantes Jr et. al. 2009], an expressive increase in the number of colors produced can be observed.

It is expected that the number of WMs, to be placed at the resulting sink nodes, will be near minimum. It is worth noticing that to find an acyclic orientation associated to (i) the minimum number of colors [Barbosa 1996], and to (ii) the maximum number of colors, are both NP-complete problems [Arantes Jr 2006].

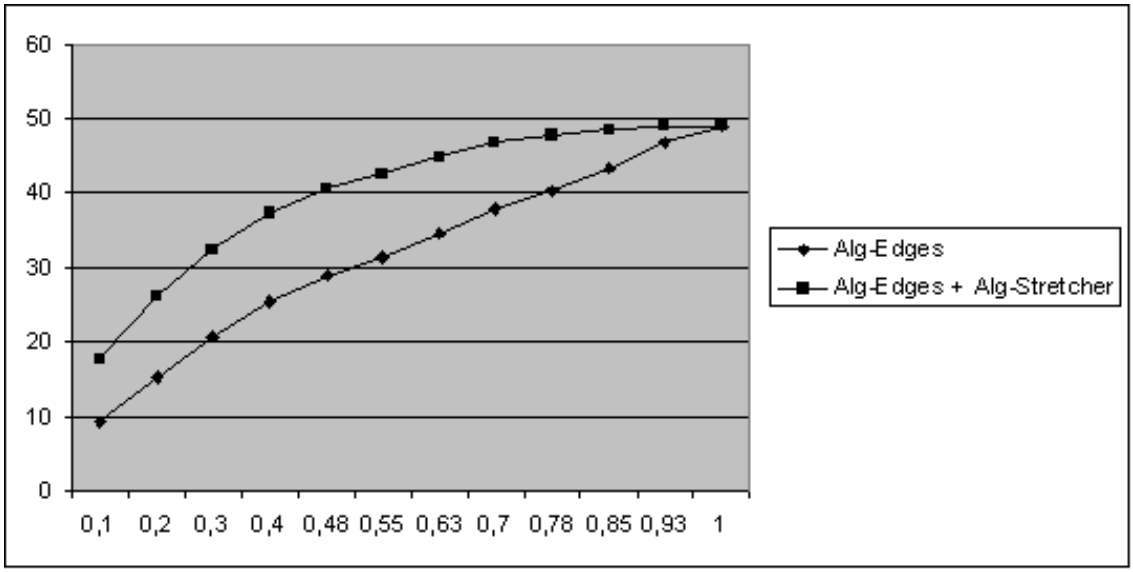


Figure 3. Numbers of colors x density

3.2 Alg-Decontamination

Let the nodes of a target directed acyclic graph G be in following three local states: **contaminated**, **clean** and **guarded**.

- **Contaminated**: the node is harmful;
- **Clean**: the node can't be contaminated again; neither of its neighbors are able to attack him;
- **Guarded**: the node contains a WM.

The following steps are taken:

WMs are placed in the *home bases* (sink nodes, $\lambda = 0$);

Sink nodes are **guarded** and all other nodes are **contaminated** (nodes have visibility 1, i.e., a WM can visualize only nodes hosting it and its immediate neighbors);

While there still exists a **contaminated** node, each node verifies its own λ. If $\lambda = 0$ and the node is **contaminated**, then it checks if it has received only one WM. If not, a leader is chosen. Once a leader is chosen, the WM that is running cleans the node, in the case of the web spam, destroy the link to the target page T and then make a decision: terminate and move to other node (a node that will become a sink) or keep the execution and make copies of itself and send to the neighbor(s) that will become a sink, i.e., send copies to its neighbors in $\lambda = 1$. The WM will not terminate its execution unless the majority of its neighbors are clean.

After sending WM copies, an ongoing WM send a message to all of its immediate neighboring nodes reversing all incident edges, producing a new acyclic orientation on G. By the end of the edge reversal from all sink nodes (in $\lambda = 0$), nodes that were in $\lambda = 1$ get into $\lambda = 0$, receiving one or more WM copies. From this point on, the process repeats itself until all contaminated nodes in the graph are extinguished.

By definition, if a node is a sink ($\lambda = 0$) then it has received at least one WM. At the moment a WM finishes its operation, by reversing all of its directed edges, such

node is no longer a sink ($\lambda > 0$). By proceeding with edge reversal over an acyclic orientation, the next orientation will be also acyclic. So, all others nodes that were previously in $\lambda > 0$ will eventually become sinks, guaranteeing the decontamination of all nodes. Notice that by keeping a WM into a node having any immediate neighbor with the majority of its neighbors **contaminated**, we avoid that an already **clean** node becomes a contaminated node once again, as indicated in [Luccio and Pagli 2007].

Figure 4 illustrates the SER-based decontamination of a 6 nodes circulant graph. In Figure 5 one can see the behavior of the asynchronous algorithm proposed by [Luccio and Pagli 2007] running over the same graph used in Figure 4.

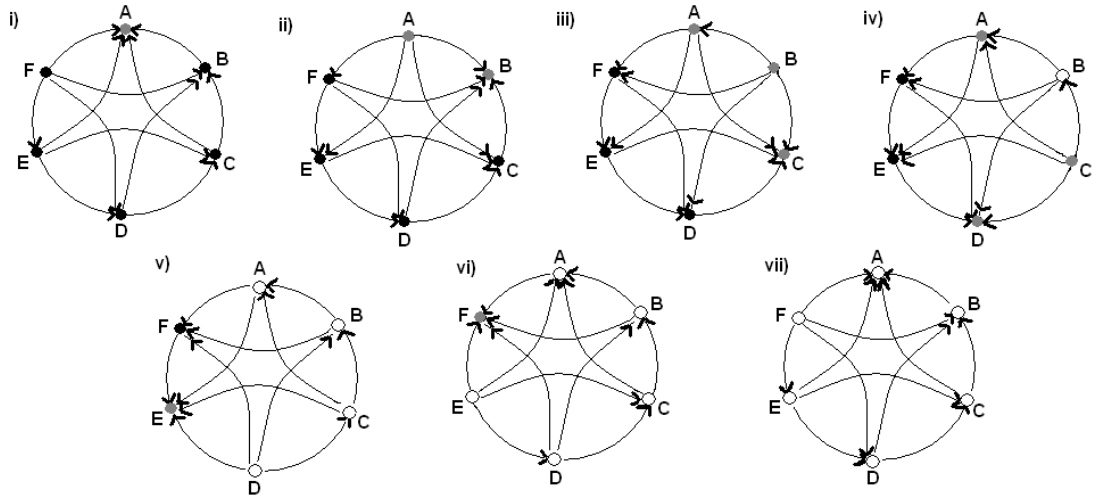


Figure 4. SER-based $C_{i,6}$ (1, 2) decontamination

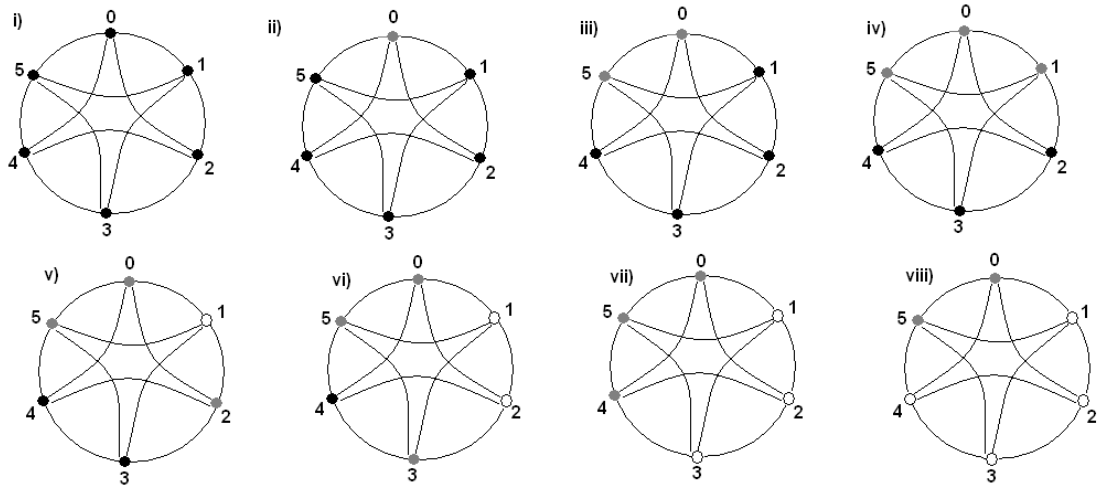


Figure 5. $C_{i,6}$ (1, 2) decontamination [Luccio and Pagli 2007]

4. Experimental Results

In order to provide a quantitative comparison between the asynchronous algorithm proposed by [Luccio and Pagli 2007] and *Alg-Decontamination*, both algorithms were applied over the same set of graph instances. The initial acyclic orientation chosen as starting point for *Alg-Decontamination* tries to reproduce the same

scenario produced by the [Luccio and Pagli 2007] algorithm, i.e., WMs "travel" only in the links of the main cycle.

Circulant graphs $C_{i,n}(L)$ with $k = 2$ ($L = \{1,2\}$) and $k = 3$ ($L = \{1,2,3\}$) are considered. In the case of $k = 2$, *Alg-Decontamination* needed three WMs, while the [Luccio and Pagli 2007] algorithm needed four WMs. In the case of $k = 3$, *Alg-Decontamination* needed four WMs, while the [Luccio and Pagli 2007] algorithm needed five WMs (all tests made with $10 \leq n \leq 10{,}000$).

In [Luccio and Pagli 2007] it is concluded that the link hops performed by WMs can be counted as the time that the decontamination takes to terminate. So a comparison of the number of link hops that were needed to decontaminate the graph is done. The number of hops needed in [Luccio and Pagli 2007] is $n - c + h$. Where n is the number of nodes as seen before, c equals to $\lfloor (k + 1)/2 \rfloor$ and h is the number of link hops needed to place the first WMs, and is represented as:

$$h = \begin{cases} 3(k^2/4 - k/2) & \text{for } k \text{ even} \\ 3(k^2/4 - k/2 + 1/4) & \text{for } k \text{ odd} \end{cases}$$

Table 1. *k*=3

	Alg-Decontamination	*[Luccio and Pagli 2007]*
10	9	11
50	49	51
1000	999	1001
3000	2999	3001
5000	4999	5001
10000	9999	10001

Concerning the number of hops, in the case of $k = 2$, *Alg-Decontamination* and the [Luccio and Pagli 2007] algorithm both needed the same number of hops, which for *Alg-Decontamination* is constant with k, and always takes $n - 1$ hops. From $k = 3$ upwards, *Alg-Decontamination* needs less link hops to decontaminate, as illustrated by Table 1. This is because it doesn't need the hops used in the [Luccio and Pagli 2007] algorithm for a verifier WM of other agents' position.

5. Conclusions

A new distributed algorithm for the decontamination of web graphs was introduced. Compared the related work, by [Luccio and Pagli 2007], the new algorithm provided the same or better figures. Moreover, while the said related work is dedicated to the class of circulant graphs, the new approach is topology independent, what could not be fully demonstrated in this paper. This suggests that Alg-Decontamination could be used in new/unseen forms of web spam. An heuristic to obtain acyclic orientations associated to the maximum number of node colors was also produced. Devising a combinatorial optimization approach to this problem is future work.

Ackknowledgements

We would like to acknowledge the Web Science Brasil project, CNPq 557.128/2009-9 and FAPERJ E-26/170028/2008 (Programa INC&T - Projeto: Instituto Brasileiro de Pesquisa em Ciência da Web).

References

Arantes Jr, G. M., França, Felipe M.G. and Martinhon, Carlos A. (2009) "Randomized generation of acyclic orientations upon anonymous distributed systems", *Journal of Parallel and Distributed Computing* Vol.69, p. 239-246.

Barbosa, V. C. (1996) "An Introduction to Distributed Algorithms". Cambridge: The MIT Press, 365 p.

Barbosa, V. C. (2000) "An Atlas of Edge-Reversal Dynamics". London: Chapman & Hall/CRC, 372 p.

França, Felipe M. G., Alves, V.C. and Granja, E. P. (2007) "Processo de Síntese e Aparelho para Temporização Assíncrona de Circuitos e Sistemas Digitais Multi-Fásicos. *Revista da Propriedade Industrial*, **1904**, INPI, PI 9703819-9.

Cassia, Ricardo F. , Alves, Vladimir C. , Bernard, Federico G.-D. and França, Felipe M.G. (2009) "Synchronous-to-asynchronous conversion of cryptographic circuits". *Journal of Circuits, Systems, and Computers*, v. 18, p. 271-282.

Luccio, F. and Pagli, L. (2007) "Web Marshals Fighting Curly Link Farm", *Proc. of FUN 2007*, LNCS, **4475**, pp. 240–248.

Flocchini, P., Nayak, A. and Schulz, A. (2005) "Cleaning an arbitrary network with mobile agents", *Proc. 2nd Int. Conference on Distributed Computing & Internet Technology*. LNCS 3816, 132-142.

Gyöngyi, Z. and Garcia-Molina, H. (2005). "Web Spam Taxonomy", In: *Proc. AIRWeb'05,* Chiba.

Donato, D., Leonardi, S., Millozzi, S. and Tsaparas, P. (2005) "Mining the inner structure of the Web graph", In *8th International Workshop on the Web and Databases (WebDB 2005)*, Baltimore, Maryland.

Flocchini, P., Nayak, A. and Schulz, A. (2007) "Decontamination of Arbitrary Networks using a Team of Mobile Agents with Limited Visibility", *Proc. of 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*.

Becchetti, L.,Castillo, C., Donato, D., Leonardi, S. and BaezaYates R. (2006) "Link-Based Characterization and Detection of Web Spam", In: *Proc. AIRWeb'06*, Seattle.

Du, Y., Shi, Y. and Zhao, X. (2006) "Using Spam Farm to Boost Page Rank", Manuscript under publication.

Lapaugh, A. (1993) "Recontamination does not help to search a graph". *Journal of the ACM*, 40 (2), 224-245,.

Barrière, L., Flocchini, P., Fraignaud, P. and Santoro, N. (2002) "Capture of an intruder by mobile agents", In: *Proc. 14th ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, Winnipeg, Canada.

Luccio, F., Pagli, L. and Santoro, N. (2006) "Nettwork Decontamination with Local Immunization", *APDCM*, 110-118.

Borie, R., Tovey, C. and Koenig, S. (2008) "Algorithms and Complexity Results for Pursuit-Evasion Problem", In: *Proceedings of the 21st international Jont Conference on Artifical intelligence* (Pasadena, California, USA, July 11 - 17, 2009). H. Kitano, Ed. International Joint Conference On Artificial Intelligence. Morgan Kaufmann Publishers, San Francisco, CA, 59-66..

Lengerke, O. , Carvalho, D. , Lima, P. M. V. , Dutra, M. S. , Mora-Camino, F. and França, F. M. G. (2008) "Controle distribuído de sistemas JOB SHOP usando escalonamento por reversão de arestas", In: *XIV Latin Ibero-American Congress on Operations Research (CLAIO 2008)*, Cartagena de Indias.

Lengerke, O. , Dutra, Max S. , França, Felipe M.G. and Tavera, Magda J.M. (2008) "Automated Guided Vehicles (AGV): Searching a Path in the Flexible Manufacturing Systems". *Journal of Konbin*, v. 8, p. 113-124.

Yang, Z. and França, Felipe M.G. (2003) "A generalised locomotion CPG architecture based on oscillatory building blocks". *Biological Cybernetics*, Heidelberg, v. 89, n. 1, p. 34-42.

Braga, R. R. , Yang, Z. and França, F. M. G. . (2008) "IMPLEMENTING AN ARTIFICIAL CENTIPEDE CPG: Integrating appendicular and axial movements of the scolopendromorph centipede", In: *International Conference on Bio-inspired Systems and Signal Processing(BIOSIGNALS)*, Funchal. Proceedings of. Setúbal : INSTICC Press, 2008. v. 2. p. 58-62.

Pinho, A. C. , Santos, A. A. , Figueiredo, D. R. and França, Felipe M.G. (2009) "Two ID-Free Distributed Distance-2 Edge Coloring Algorithms for WSNs", In: *8th International IFIP-TC 6 Networking Conference (NETWORKING 2009)*, Aachen. LNCS. Berlin / Heidelberg : Springer, 2009. v. 5550. p. 919-930.

Arantes Jr., G. M. (2006) "Trilhas, Otimização de Concorrência e Inicialização Probabilística em Sistemas sob Reversão de Arestas" Tese (Doutorado em Engenharia de Sistemas e Computação) - Universidade Federal do Rio de Janeiro.