Improving Distributed Resource Search through a Statistical Methodology of Topological Feature Selection

Claudia Gómez Santillán
Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, IPN, México
Instituto Tecnológico de Ciudad Madero, Cd. Madero, México
Email: cgomezsa@ipn.mx, cggs71@hotmail.com

Laura Cruz-Reyes, Eustorgio Meza, Tania Turrubiates López, Marco A. Aguirre Lam, and Elisa Schaeffer
Instituto Tecnológico de Ciudad Madero, Cd. Madero, México, Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, IPN, México, Instituto Tecnológico de Alamo Temapache, Veracruz, México, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, México
Email: lcruzreyes@prodigy.net.mx, emeza@ipn.edu.mx, tania_251179@acm.org, marco@marcoaguirre.com.mx, elisa@yalma.fime.uanl.mx

***Abstract*—The Internet is considered a complex network for its size, interconnectivity and rules that govern are dynamic, because of constantly evolve. For this reason the search of distributed resources shared by users and online communities is a complex task that needs efficient search method. The goal of this work is to improve the performance of distributed search of information, through analysis of the topological features. In this paper we described a statistical methodology to select a set of topologic metrics that allow to locally distinguish the type of complex network. In this way we use the metrics to guide the search towards nodes with better connectivity. In addition we present an algorithm for distributed search of information, enriched with the selected topological metric. The results show that including the topological metric in the Neighboring-Ant Search algorithm improves its performance 50% in terms of the number of hops needed to locate a set of resources. The methodology described provides a better understanding of why the features were selected and aids to explain how this metric impacts in the search process.**

***Index Terms*— Internet, search process, query routing, random walk, ant colony system, scale free, topology, experiment designs, statistical analysis, metrics**

I. INTRODUCTION

Complex systems can be modeled by means of complex networks, because of have a non-trivial topological structure. These features have motivated the study of topological features of real-world networks such as the Internet. Knowledge on such features can be used to optimize the performance of processes carried out on the Internet, for example: the search of distributed resources, traffic management, and design of routing queries [1], among others. The main goal is to help the users to find the information that they request with a reasonable processing time and with a higher quality of the information obtained [2].

Over the past years, new communication models have emerged in the Internet that manage information in a distributed manner and offer significant advantages over centralized information management systems. These systems are known as *peer to peer* networks (P2P). In a P2P network, a set of nodes form connections to offer their resources to the other nodes within the network. The P2P systems, together with the underlying communication network (typically the Internet), form a complex system that requires autonomous operation through mechanisms of intelligent search [1].

Until now, a great number of topologic metrics have been developed to characterize the complex networks, but the majority of these metrics are global. This implicates a great computational effort (processing time and memory). For this reason, it is necessary to identify a topological metric that locally allows to obtain information about the type of network. In this way the distributed search process would take advantage of the topology. This point requires sufficient empirical evidence that supports the use of a topological metric to locally identify the type of network. This raises important questions: what topological metric to select in order to locally identify the types of complex networks? Does the topologic metric allow to improve the performance of the distributed search process? If so, how much the performance of the distributed search process is improved? Why the topological metric can identify the type of network?

In this paper, a methodology based in statistical analysis is described. The goal of this methodology is to identify, by the means of an experimental design and a series of statistical tests, a set of topologic metrics that allow to locally recognize the type of a complex network. This minimum set is used to analyze the performance of a distributed search algorithm for textual information,

enriched with a topological metric to characterize local topology. The study of such distributed algorithms is important as the quality of the retrieved information as well as the time necessary for its retrieval are key factors in the performance of a P2P system.

II. Theoric Frame

A. Peer-to-Peer Networks Modeled as Complex Networks

A *system* is a set of interrelated components that seeks a common goal. Any system that can be understood as a set of components whose connections follow a certain rule can be modeled as a network: each component is represented by a node of the network and all existing interactions are captured by the connections of the network [3].

Complex systems are those systems that have a very large number of components and the connections among the components may evolve over time and the roles of the components may vary. In many studies, complex systems are modeled as networks, giving rise to the concept of complex networks [3].

A *P2P network* is a distributed system where all the nodes are equal in terms of functionality and tasks performed in the P2P system [4]. The objective of a P2P network is to share resources such as information (documents, music, videos), hardware resources (computational capacity, memory), or peripheral devices (printers, cameras).

Such structure formed by pairs of connected nodes can be modeled as the edges of a dynamic network, where edges and nodes may appear and disappear at any time. Hence the structure of a P2P network can be modeled as a complex network [1, 5]. One of the main motivations for modeling systems as complex networks is the flexibility and generality of the abstract representation that allows handling properties such as dynamic topology in a natural way [6].

B. Information Search

The problem of locating textual information in a P2P network over the Internet is known as *semantic query routing* (SQR), where the goal is to determine the shortest paths from a node that issues a query to those nodes that can appropriately respond to the query (by providing the requested information) [1]. The query traverses the network moving from the initiating node to a neighboring node and then to a neighbor of a neighbor and so forth until locating the requested resource (or giving up in its absence).

The challenge lies in the design of algorithms to traverse the Internet in search of resources – modeled as a complex network – in an intelligent and autonomous manner. In order to reach this goal, the algorithms proposed for this problem include the selection of the next node to visit, using information of near-by nodes of the current node, that is, information on the *local topology* of the current node.

C. Random Walk

The *random-walk* (RW) search algorithm is a blind search technique where the nodes of the network possess no information on the location or contents of the requested resource unless the resource resides in the node itself [8]. Let G be a graph that models the network and v a vertex in G. A *T-hop random walk* from v in G is a sequence of dependent random variables X_0 ,..., X_T defined as follows: $X_0 = v$ with probability 1 and for each $i = 1,... T,$ the value for X_i is selected uniformly at random among the vertices $\Gamma(X_{i-1})$, that is, among the neighbors of the vertice of the preceding step. Simply put, a random walk begins at a certain vertex and on each step, moves to a neighbor of the current vertex, until it arrives to a vertex that meets to goal. In our network, that would be a vertex that represents a node that contains the requested resource [7].

D. Neighboring-Ant Search

In the area of classification, feature selection has great benefits such as improving the performance of classification procedures and constructing simple and comprehensible classification models. These are achieved eliminating irrelevant and redundant features that may introduce noise that could affect the efficiency of the procedure. The goal of feature selection is to choose a minimum subset of features that can discriminate efficiently among different classes. This minimum subset is known as the *optimal* subset [24]. The majority of the feature selection methods involve the search in the feature space to predict the best class and the evaluation of the features to measure the fitness of a subset [25].

E. Experimental Design

Understanding a particular system or process demands observation, modeling, and experimentation. The experimentation aims to generalize away from context specific measurements and to build insight into fundamental structures and properties of a system. In this way the experimentation provides knowledge of the domain of interest [11, 12, 13].

Learning involves the encapsulation of knowledge, checking that the knowledge is correct, and evolving that knowledge over time. The experimental paradigm is used in many fields, including physics, medicine, and manufacturing. Like other sciences, many disciplines within computer science likewise require an empirical paradigm [11, 14]. Since the experimental subject and the research questions are somewhat unusual compared to

other problem domains, much more work is needed to identify the statistical and data analysis tools most appropriate to these types of problems [12, 13, 14].

F. Degree Disperion Coefficient

The DDC measures the differences between the degree of a vertex and the degrees of its neighbors. A node i is said to be a *neighbor* of a node j if they are connected in the network. The *degree* k_i of a node i is the number of neighbors it has. The DDC of node i is defined in equation (1), $\sigma(i)$ is the degree variation among i and its neighbors and $\mu(i)$ is their average degree [26].

$$DDC(i) = \frac{\sigma(i)}{\mu(i)};$$

$$\text{where } \sigma(i) = \sqrt{\frac{\sum_{j\in\Gamma(i)} \left[k_j - \mu_i\right]^2 + \left[k_i - \mu_i\right]^2}{k_i + 1}}; \qquad (1)$$

$$\text{and } \mu(i) = \frac{\sum_{j\in\Gamma(i)} \left[k_j\right] + k_i}{k_i + 1}$$

III. Related Works

Existing methods for classifying real-world networks using topological features [9, 10, 15] do not provide a detailed statistical analysis to determine if all of the features used are necessary or optimal to efficiently discriminate among types of networks. Costa *et. al.* [6] used statistical techniques to identify the type of a network with unknown nature. The results show that the type of network assigned to the networks, varies according to the topological features selected, and that excessive number of features can compromise the quality of the classification.

In the above mentioned methods the topological features used are *global*: computing each feature requires processing the entire network. This involves a great computational effort with respect of both time and memory.

One of the problems of interest is the semantic query routing on the Internet. The most relevant works in this area address this problem using ant-colony algorithms [1, 22, 23]. The principal difference of the present work with existing methods is the incorporation of a strategy that takes advantage of the environment where the search takes place, in terms of a local structural metric is. The structural metric was selected through a statistical methodology described in Section 3.

IV. Statistical Methodology

The main goal of the statistical methodology proposed is to identify which topological features are relevant and non-redundant. Let us first define relevancy feature. Let us suppose there are k different populations of networks, from which topological features are extracted. Three possible cases can be identified, c.f Figure 1:

- ***Strong Relevance:*** the differences among the k types of networks are sufficient to be able to classify a new network based on the features measured and the populations do not overlap.
- ***Weak Relevance:*** the differences among the k types are strong but insufficient to perfectly classify a new network due to population overlap.
- ***Irrelevancy:*** the differences among the k types of networks are insufficient for classification.

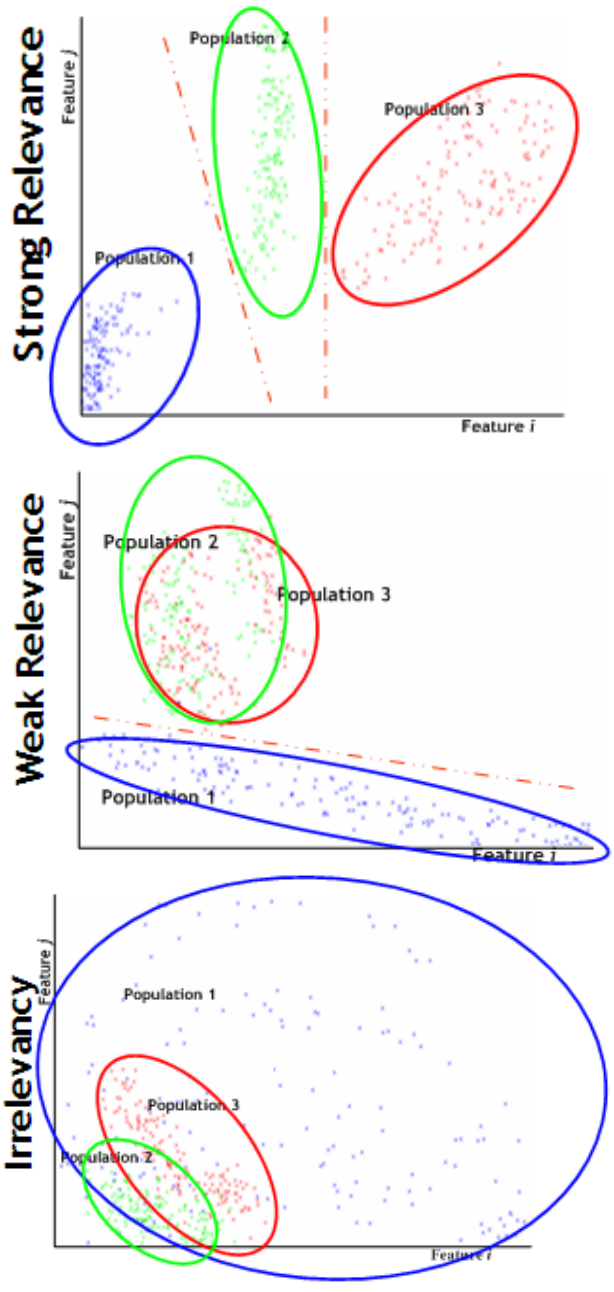


Figure 1. Three possible cases of relevancy feature.

The statistical methodology proposed to select relevant and non-redundant features has four basic steps; *c.f.* Figure 2:

1. **Identify relevant and irrelevant features:** This step consist in determining which features differ significantly according to the type of network, regardless of the number of nodes in the network. The features with different means are considered as relevant features whereas the features with equal means are irrelevant. The experimental design used to discriminate between relevant and irrelevant features is described by Cruz *et al.* [16] and Turrubiates *et al.* [17]
2. **Identify features with strong relevance and weak relevance:** In this step the relevant features are analyzed by means of a multiple comparison method to determine which of them are strong relevant features and which are weak relevant features.
3. **Redundancy elimination:** The goal in this step is to eliminate the weak relevant features correlated with the strong relevant features in such a way that the set of selected features contain the majority of the strong features and some of the weak features.
4. **Identify the minimal set:** Combinations of selected features are made and a discriminant analysis is carried out to determine the combination with the lowest number of features that produces the best performance in the discriminant analysis. This combination is called the *minimal set*.

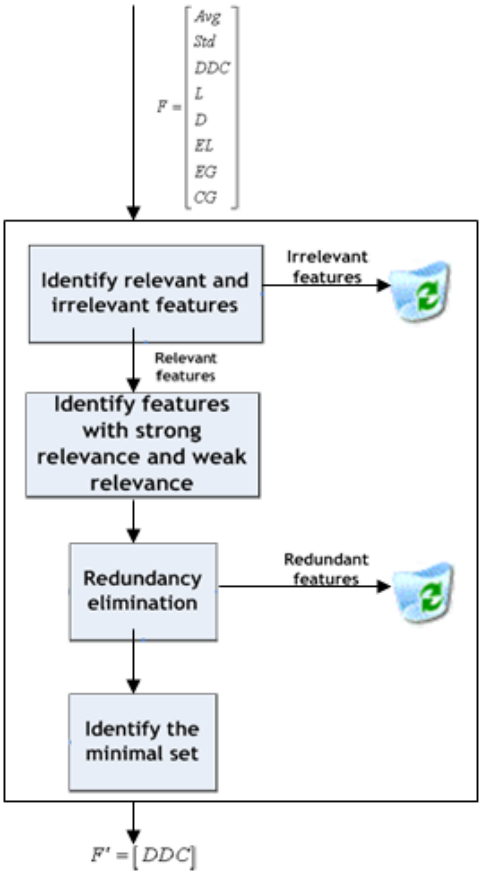


Figure 2. Steps of feature selection

V. Experiments

A. Statistical Methodology

The topological features analyzed were: average degree *Avg*, standard deviation of the degree *Std*, clustering coefficient *CG*, global efficiency E_{glob}, local efficiency E_{loc}, shortest path length *L*, diameter *D* and the degree dispersion coefficient *DDC*. A detailed description of these features could be found in Costa *et. al.* [6]. The methodology considers that the features could be affected by the type of network and the number of nodes; we generated three kinds of complex networks with different sizes [18].

B. Distributed Resource Search

We generated complex networks with the scale-free network method of Barabási et al. [20], where nodes are added one at a time with a fixed number of connections each. The newly-arriving node chooses preferentially at random among the existing nodes to which to connect, giving preference to high-degree nodes. The resulting network has a small number of highly connected nodes while the majority of the nodes have degree close to average degree. The network size was set to 1,024 nodes.

The studied algorithms were the proposed Neighboring-Ant Search (NAS) and the random-walk (RW) algorithm that serves as a base case for comparison. We experimented on two versions of both algorithms: with and without the DDC. Only one ant was used per query and the time-to-live of the ants was set to 25 hops. The number of results (hits) needed to satisfy the query was set to five.

The time steps of the experiments were of 100 ms and the simulations were ran for 10,000 steps. During each time step, each node has a probability of 0.1 to launch a query. The "topics" of the resources were modeled as integer values from zero to 1,024, generated using the uniform-distribution generator of Repast [21]. Each node was assigned ten resources with possibly repeated topics. The queries were generated to search a topic uniformly at random from 0 to 1,024. The experiments were carried out on a workstation with an Intel Xeon a 3GHz processor with 4GB of RAM.

C. Random Walk

Optionally, one can include the DDC function into the random-walk algorithm. A simple modification to include structural preferentiality is to choose uniformly at random two neighbors, calculate their DDC values, and move on to the neighbor with higher DDC.

D. Neighboring-Ant Search

The NAS algorithm has two objectives: directing the

TABLE I. NAS ALGORITHM PSEUDO-CODE

```
01 for each query
02  repeat while the forward ant is active
03  if Hits < maxResults and TTL > 0        // Phase 1
04   if the neighbor from edge s_k has results
05    append s_k to Path_k
06    TTL_k = TTL_k -1
07    globalUpdate  // backward ant
08   else                                   // Phase 2
09    s_k = apply the transition rule
10    if path does not exist or node was visited,
11     remove the last node from Path_k
12    else,
13     append s_k to Path_k
14     TTL_k = TTL_k -1
15     localUpdate
16    endif
17  endif
18 else
19   Kill the forward ant
20 endif
21 endif
```

queries towards the nodes that have good connectivity using the DDC while minimizing the number of hops needed to respond to queries. This latter objective is achieved by a function called *importance of hops* that is the inverse of the sum of all connections traversed during the search. The number of hops is the lifetime of the ant

TABLE II.
STATISTICAL TOOLS USED IN THE PROPOSED METHODOLOGY

Step	Statistical test	Results
Identify relevant and irrelevant features	Experimental Design: Two factor mixed factorial Statistical Test: MANOVA, Residuals Analysis, Interaction Plots.	✓ Relevant features: *Std*, *DDC*(*G*), *L*(*G*), *D*(*G*), E_{loc}, E_{glob} × Irrelevant features: *Avg*, *CG*(*G*)
Identify features with strong relevance and weak relevance	Multiple comparison using the Tukey test	Strong relevant features: *Std*,, *DDC*(*G*), *L*(*G*), *D*(*G*) Weak relevant features: E_{loc}, E_{glob}
Redundancy elimination	Correlation Analysis	✓ Not redundant features: *Std*,, *DDC*(*G*), *L*(*G*), E_{loc} × Redundant features: *D*(*G*), E_{glob}
Identify the minimal set	Discriminant Analysis	Minimal set: *DDC*(*G*).

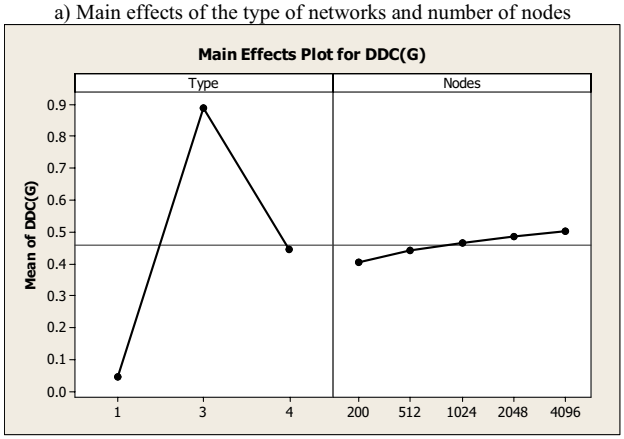
represented by *TTL* and the maximum value is set at 25 hops. A pseudo-code for the algorithm is given in Table 1 [19].

VI. RESULTS

In Table 2, the statistical tests used in each step of the methodology are described together with the obtained results, *c.f* [16, 18].

The subset obtained in the steps that correspond with

a) Main effects of the type of networks and number of nodes



b) Interaction of the type of networks and number of nodes

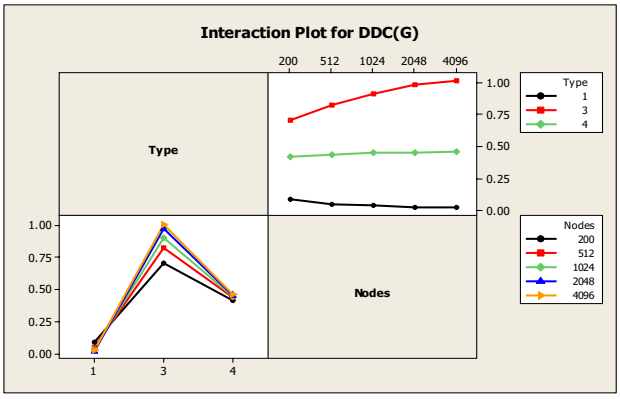


Figure 3. Factor effects for the *DDC* (*G*) feature.

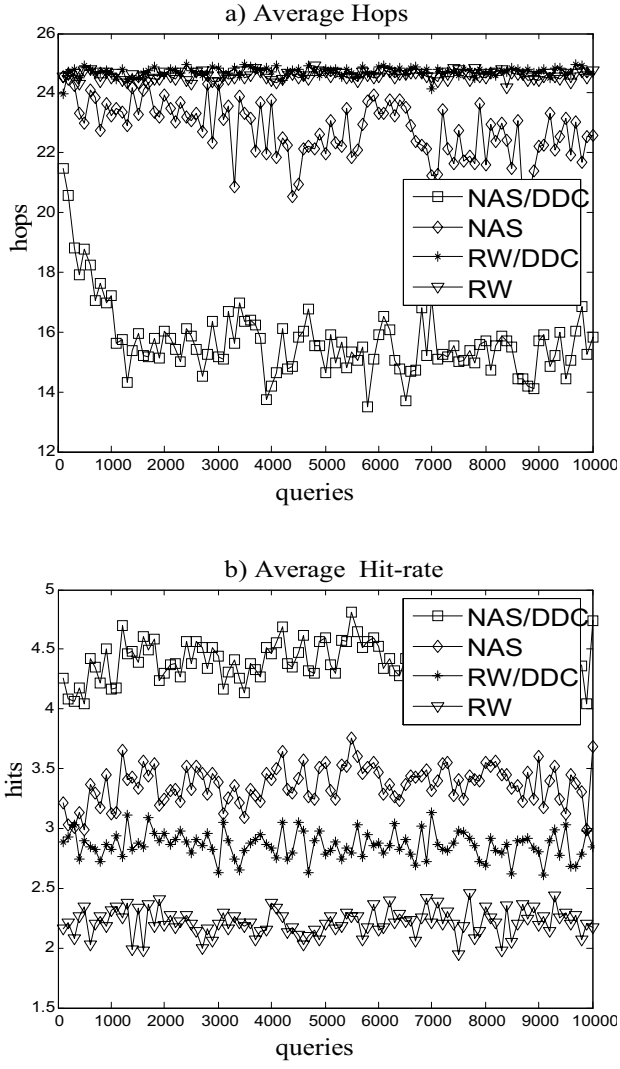
the search in the feature space is formed by the standard deviation of the degree $\sigma_{(k)}$, local efficiency E_{loc}, shortest path length *L*, and the degree dispersion coefficient *DDC*. The evaluation of this subset identified the *DDC* as the feature that by its own has a best performance in the discriminant analysis.

The *DDC* was selected (*c.f* Figure. 3) as the relevant and not redundant feature that can discriminate efficiently among the three types of complex networks. The analysis of the topological features selected shows that the type of network has a greater effect on the results than the number of nodes in the network. This is important in locally identifying the type of a network; this feature captures sufficient information about the types of networks [18].

After having selected the DDC as the relevant and not redundant feature, was included into the two search algorithms, obtaining the following results. In Figure 4(a), it can be seen that in scale-free networks, the

number of hops used by the RW algorithm is very close to the TTL of 25 hops, with or without the DDC. For the NAS algorithm without DDC, approximately 19 hops are needed. Incorporating DDC into NAS reduces the number of hops even further down to 10.

Figure 4(b) shows the average hit-count. On scale-free topologies, the RW algorithm obtains 2 – 2.5 results per query without using the DDC, and with DDC, the number of hits rises to 2.5 – 3. In the same networks, the NAS algorithm obtains 3 – 3.5 hits per query without the DCC, and 4 – 4.5 hits when DCC is used.



These observations confirm the intuition that the DDC in the presence of a scale-free distribution allow a significant improvement to the search performance, which also implies that the NAS algorithm outperforms in such topologies the existing methods that do not incorporate local structural information.

VII. Conclusions And Future Work

In this work a statistical methodology was defined and developed to identify the minimal set of topological features that allows to discriminate among three different types of complex networks. The use of this methodology allows us to justify why the features were selected and provide information of the influence of the type of network and the number of nodes in the prediction power of the selected features.

The result of applying the methodology to a set of eight topological functions resulted on minimal set containing the DDC metric [18]. Subsequently, this metric was used in semantic query routing algorithms. We observe that upon including DDC in the algorithm NAS, the hop count decreases by 50% and the hit count is improved by 15%. The random-walk algorithm used as a comparison gains no advantage of DDC in terms of the hop count, and a benefits very little in terms of hit count (3% improvement).

As future work, we plan to study more profoundly the impact of the metrics employed in the learning curve of ant-colony algorithms as well as the effect on the performance measures of hop and hit counts. We also contemplate using more than one ant per query to parallelize the algorithm in hopes of improved performance.

References

[1] Michlmayr E.: *Ant Algorithms for Self-Organization in Social Networks*. Ph.D. Thesis, Vienna University of Technology, Vienna, Austria, 2007.

[2] Arenas A., Danon, L., Díaz-Guilera, A., and Guimera, R.: Search and Congestion in Complex Networks. In *Statistical Mechanics of Complex Networks: XVIII SITGES Conference on Statistical Mechanics*, Lecture Notes in Physics, vol. 625, 175 - 194. Springer Verlag, 2003.

[3] Amaral, L. A. N., and Ottino, J. M.: *Complex Systems and Networks: Challenges and Opportunities for Chemical and Biological Engineers.* Chemical Engineering Scientist (59): 1653–1666, 2004.

[4] Androutsellis-Theotokis, S., and Spinellis, D.: *A Survey of Peer-to-Peer Content Distribution Technologies.* ACM Computing Surveys, 36(4): 335–371, 2004.

[5] Ortega, R.: *Estudio de las Propiedades Topológicas en Redes Complejas con Diferente Distribución de Grado y su aplicación en la búsqueda de recursos distribuidos*. Technical Report, Centro de Investigación en Ciencia Aplicada y Tecnología Aplicada, Altamira, Mexico, 2005.

[6] F. Costa, L., Rodriguez, F. A., Travieso, G., and Villas Boas, P. R.: *Characterization of Complex Networks: A survey of measurements*, Advances in Physics, 56(1): 167-242, 2007.

[7] Arora, S., and Barak, B.: *Complexity Theory: A Modern Approach. Book in preparation*, 2007.

[8] Dorigo, M., and Gambardella, L. M.: *Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem*. IEEE Transactions on Evolutionary Computation, 1(1): 53-66, 1997.

[9] Airoldi E.M., and Carley K.M.: *Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings.* ACM SIGKDD Explorations Newsletter 7:13-22, 2005
[10] Middendorf M., Ziv E., Carter A., Hom J., Koytcheff R., Levovitz C., Woods G., Chen L., and Wiggins C.: *Discriminative Topological Features Reveal Biological Network Mechanisms.* BMC Bioinformatics 5:181, 2004.
[11] Basil V.R., and Zelkowitz M,V.: *Empirical Studies to Build a Science of Computer Science*. Communications of the ACM 50:33-37, 2007.
[12] McGeoch C.C.: *Experimental Algorithms.* Communications of the ACM 50:27-31, 2007.
[13] Montgomery D.C.: *Design and Analysis of Experiments.* John Wiley & Sons. New York, 2001.
[14] Hooker J.N.: Needed: *An Empirical Science of Algorithms.* Operations Research 42:201-212, 1994.
[15] Ali W., Mondragón R.J., and Alavi F.: *Extraction of topological features from communication network topological patterns using self-organizing feature maps.* arXiv:cs/0404042v2, 2004.
[16] Cruz R.L., Meza C. E., Turrubiates L.T., Gomez S.C., and Ortega I.R.: *Experimental Design for Selection of Characterization Functions that Allows Discriminate Among Random, Scale Free and Exponential Networks*. Polish Journal of Environmental Studies 16:67-71, 2007.
[17] Turrubiates L.T., Gómez S.C., Cruz R.L., Ortega I.R., and Meza C. E.: *Diseño Experimental para Selección de Funciones de Caracterización que Permitan Discriminar entre Redes Aleatorias, de Escala Libre, y Exponenciales.* 14th International Congress on Computer Science Research CIICC'07. 161-171, 2007.
[18] Turrubiates L.T.: *Clasificación de Redes Complejas usando Funciones de Caracterización que Permitan Discriminar entre Redes Aleatorias, Power-Law y Exponenciales.* M.Sc. Thesis, IT Cd. Madero, Madero, 2007.
[19] Cruz-Reyes L., Gómez S.C., Aguirre M.A.L., Schaeffer E.S., Turrubiates L.T., and Ortega I.R.: *NAS Algorithm for Semantic Query Routing Systems in Complex Networks*, accepted in DCAI'08.
[20] Barabási A.L., Albert R., and Jeong H.: *Mean-Field theory for Scale-free Random Networks*. *Physica A,* 272: 173–189, 1999.
[21] ROAD: *Repast Organization for Architecture and Design*. Repast Home Page. Chicago, IL, USA. http://repast.sourceforge.net, 2005.
[22] Yang K.H., Wu C., and Ho J.M.: *AntSearch: An Ant Search Algorithm in Unstructured Peer-to-Peer Networks*. IEICE Transactions on Communications, 89(9): 2300-2308, 2006.
[23] Babaoglu O., Meling H., and Montesor A.: *Anthill: A Framework for the Development of Agent-Based Peer-to-Peer Systems*. In Proceedings of the 22nd International Conference on Distributed Computer Systems (ICDCS 02). IEEE, 2002.
[24] Yu, L., Liu, H.: *Efficient Feature Selection via Analysis of Relevance and Redundacy.* Journal of Machine Learnig Research 5. 1205 – 1224, 2004
[25] Witten, H.I., Frank, E.: Data Mining. *Practical Machine Learning Tools and Techniques.* Second Edition. Elsevier, 2005.
[26] Ortega, R., Meza, E., Gómez, G., Cruz, L., Turrubiates, T.: *Impact of Dynamic Growing on the Interner Degree Distribution*. In Frontiers of High Performance Computing and Networking ISPA 2007 Workshops, Lecture Notes in Computer Science. vol. 4743.Springer Verlag. 119-122, 2007.

Claudia Gómez S. is a doctoral student at National Polytechnic Institute, Mexico. She received her MS degree in Computer Science from the Leon Institute of Technology, Mexico, in 2000. Her research interests are optimization Techniques, complex network and autonomous agents.

Laura Cruz-Reyes was born in Mexico in 1959. She received the PhD (Computer Science) degree from National Center of Research and Technological Development, Mexico, in 2004. She is a professor at Madero City Institute of Technology, Mexico. Her research interests include optimization techniques, complex networks, autonomous agents and algorithm performance explanation.

Eustorgio Meza received the PhD (Oceanic Engineering) deegree from Texas A&M University, College Station, U.S.A. He received the MS degree in Computer Science (AI) from Institute of Technology and Advanced Studies of Monterrey, Mexico. He is a Professor at Research Center in Applied Science and Advanced Technology from National Polytechnic Institute. His research interests are oceanology, complex Network.

Tania Turrubiates López is a doctoral student in Autonomous University of Nuevo Leon. She obtained her MS degree in Computer Science from Madero City Institute of Technology, Mexico in 2007. Her research interests are optimization techniques and complex network.

Marco Aguirre Lam. is a MS student in Computer Science at Madero City Institute of Technology, Mexico. Her research interests are optimization technique, complex network and autonomous agents.

Elisa Schaeffer was born in Finland in 1976. She received the PhD (Science in Technology) degree from Helsinki University of Technology, Finland, in 2006. She is a teaching researcher at the Graduate Program in Systems Engineering (PISIS), at the Faculty of Mechanical and Electrical Engineering (FIME), Universidad Autonoma de Nuevo Leon, Mexico. Her research interests include nonuniform networks, graph clustering and optimization of network operation.