U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Dep. de Ciência de Computadores

MAPi PhD Thesis Proposal

On the Characterization and Comparison of Complex Networks

Fernando Silva (fds@dcc.fc.up.pt)
CRACS & INESC-Porto LA - FCUP

Note: This proposal is intended for the student Sarvenaz Choobdar who is being sponsored by a research grant from INESC-Porto LA.

Motivation A wide variety of real life structures can be intuitively represented by complex networks [1]. Mining interesting features from these networks is a very important task with an inherent multidisciplinary impact [2]. Complex networks are special kinds of graphs and thus, much of the theorectical work and graph concepts hold for these networks. Metrics such as average degree or the clustering coefficient can, in many cases, provide relevant information that helps one to characterize and distinguish different types of networks [3].

One very important associated concept is that of network motifs [4]. These are recurrent significant patterns that are overrepresented, that is, appear in much higher frequency than it would be expected in a similar network. Although relatively recent (the term was coined in 2002), motifs were shown to be a very powerful and useful general idea, with published applications in a multitude of domains, ranging from networks in protein-protein interaction, gene transcriptional regulation, brain, food webs, electronic circuits and even software. Finding network motifs is a computationally hard problem (due to graph isomorphism). Our group has contributed in the past two years in developing efficient data structures, sequential and parallel algorithms, for motif detection in complex networks. We surveyed all the main strategies [14], contributed with an efficient data structure, g-tries [16], efficient approximation heuristics [17] and we contributed with parallel algorithms for subgraph enumeration and motif finding [15,18].

A survey of network comparison methodologies can be seen on [5]. Research is still on its early development stages and motif use on this realm is still scarce. Pruzlj [6] uses graphlet distributions, but utilizes a pre-defined static set of small subgraphs. Costa et al [7] show how single nodes can be compared to each other based on quantitative measurements but do not use motif related metrics and use information compression that can cause some loss of information.

Although the analysis of properties of static graphs has been the focus of many studies, discovering the trends over time as networks evolve over long periods is still a very recent area. Some initial studies have already been made [8,9], but work on understanding the role of network motifs as networks evolve over time is still very scarce. An adaptation of the concept has already been suggested in the form of trend motifs [10], and some other local topology related measurements have been used, such as frequent subgraphs over time [11].

Goals Our main goals are threefold, with emphasis on using network motifs or associated concepts for the following objectives:

- Develop methodologies to compare, distinguish and characterize different complex networks;
- Develop methodologies to compare, distinguish and characterize individual network nodes;
- Develop methodologies to characterize and analyze networks over time.

We need to be able to answer questions like: how similar are two different networks? How different are two networks? What makes them different and which trait better distinguishes it? Given two sets of networks, is another given unclassified network more close to one or other set? How closely related are two different nodes in the same network? And in different networks? Can we identify and measure the importance of a node? What trends develop, remain and change as networks evolve over time?

Research description The research within this proposal is organized in two main lines that are related. It may be that as the work progresses, more emphasis will be given to just one of these lines. The lines are:

Metrics for characterization and comparison: Study, implement and extend known (and new) motif metrics, like the motif frequency spectra [12] or node motif contribution [13] in order to verify both their scalability and usefulness for the larger scale allowed by our methodology, effectively developing large motif fingerprints. We plan to compare different networks by using distance measures both from the previously referred motif fingerprints and by adapting related methodology. In particular, we plan to extend the graphlet degree distribution [6], which is based on a static pre-defined set of 73 different subgraph and their respective frequencies, and obtain an algorithm that dynamically chooses the motifs to use as 'graphlets' to include in the comparison. We also plan to measure the centrality of nodes, by computing high-dimensional feature vectors of nodes that include both motif metrics and other local features. Such as in [7], we plan to reduce the dimension of the feature vector in order to have more directly comparable quantitatives that would allow us to discover singular nodes.

Networks over time: Study the behavior and influence of motifs in the dynamics of time evolving networks. We plan to analyze several motif metrics in natural and synthetic networks as they change over time and we intend to discover how motif fingerprints evolve. Our goal is to both characterize the time evolution and be able to make informed predictions on future network states. We intend to construct sequences of data points from quantitative metrics associated both with the global network and single nodes, and apply statistical time series analysis, looking at the problem from a more classical mathematical perspective. We will also try to understand if slight changes in the network motif concept, such as 'trend motifs' [10] are more suitable for this kind of analysis.

From start we leave our lines of research still open, but we expect to narrow their scope early to achieve our main goal that is to understand the evolution of networks over time, such as in [9], and how network motifs both influence and are influenced by it. Of particular interest to us are the evolution and characterization of networks with fixed nodes but evolving connections among them, as is expected in brain networks.

The work plan schedule takes this in consideration:

- (1st year) Study main concepts on graph theory and algorithms and time series analysis. Research the state of the art on algorithms for motif discovery, network metrics and time

evolving networks, specially those that have fixed nodes. Write, submit and defend a thesis proposal within the MAPi program.

- (2nd/3rd year) Define a relevant set of metrics for network characterization and motif fingerprinting. Do a case study application on a real data set and assess positive and negative quality of the metrics used. Use these metrics on evolving networks and derive new time series for a different angle of analysis.
- (4th year) Work on applications and quality of the metrics proposed. Thesis writing.

The student will be encourage to revise the state of the art as the work progresses given that the research in this area is very prolific.

References

1. R. Albert and A. L. Barabasi, Statistical mechanics of complex networks, Reviews of Modern Physics, vol. 74, no. 1, 2002.
2. L. da F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. V. Boas, L. Antiqueira, M. P. Viana, and L. E. C. da Rocha, Analyzing and modeling real-world phenomena with complex networks: A survey of applications, ArXiv e-prints, vol. 0711, no. 3199, 2007.
3. L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas. Characterization of complex networks: A survey of measurements. Advances In Phyisics, vol. 56, p. 167, 2007.
4. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U.Alon, Network Motifs: Simple Building Blocks of Complex Networks. Science, 298:824-827, 2002.
5. R. Sharan and T. Ideker, Modeling cellular machinery through biological network comparison. Nature biotechnology, vol. 24, no.4,pp. 427-433, April 2006.
6. N. Przulj, Biological network comparison using graphlet degree distribution, Bioinformatics, vol. 23, no. 2, pp. e177-183,January 2007.
7. L. da F. Costa, F. A. Rodrigues, C. C. Hilgetag and M. Kaiser. Beyond the average: Detecting global singular nodes from local features in complex networks. EPL (Europhysics Letters), vol. 87, no.1, pp 18008.
8. Prasanna Desikan and Jaideep Srivastava. Mining temporally changing web usage graphs. WebKDD, 2004.
9. Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining, pages 177-187, New York, NY, USA, 2005. ACM.
10. R. Jin, S. McCallen and E.d Almaas. Trend Motif: A Graph Mining Approach for Analysis of Dynamic Complex Networks. 7th IEEE International Conference on Data Mining, pp. 541-461, 2007.
11. M. Berlingerio, F. Bonchi, B. Bringmann and A. Gionis. Mining graph evolution rules. Machine Learning and Knowledge Discovery in Databases, Springer, pp. 115–130, 2009.

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Dep. de Ciência de Computadores

12. O. Sporns O, C.J. Honey and R. Kötter. Identification and Classification of Hubs in Brain Networks. PLoS ONE 2(10): e1049, 2007

13. F. Schreiber and H. Schwöbbermeyer. MAVisto: a tool for the exploration of network motifs. Bioinformatics, 21, 3572-3574, 2005.

14. P. Ribeiro, F. Silva and M. Kaiser. Strategies for Network Motifs Discovery. 5th IEEE International Conference on e-Science, Oxford, UK, pp 80-87, IEEE CS Press, December, 2009.

15. P. Ribeiro, F. Silva and L. Lopes. Parallel Calculation of Subgraph Census in Biological Networks. Proceedings of the 1st International Conference on Bioinformatics, Valencia, Spain, pp 56-65, INSTICC, January, 2010.

16. P. Ribeiro and F. Silva. G-Tries: an efficient data structure for discovering network motifs. Proceedings of the ACM 25th Symposium On Applied Computing - Bioinformatics Track, Sierre, Switzerland, pp 1559-1566, March, 2010.

17. P. Ribeiro and F. Silva. Efficient Subgraph Frequency Estimation with G-Tries. To appear in 10th Workshop on Algorithms in Bioinformatics, Springer LNBI, September, 2010.

18. P. Ribeiro, F. Silva and L. Lopes. Efficient Parallel Subgraph Counting using G-Tries. To appear in the IEEE International Conference on Cluster Computing, IEEE CS Press, September, 2010.

Porto, 6th December 2010

Fernando Manuel Augusto da Silva

Rua do Campo Alegre 1021, 4169-007 Porto Tel: 220402900 Fax: 220402950 URL: www.dcc.fc.up.pt