PhD thesis proposal - 2015

Lossless graph compression to assist biological network understanding

Contact: Jacques.Nicolas[AT]irisa.fr
Project-Team: Dyliss (Bioinformatics)
Supervisor: Jacques NICOLAS, DR INRIA
Possible funding and Starting Date: MENRT: October 2015

Context: The production of large graphs has been boosted by internet applications (web, social networks) as well as technical (communication networks, power networks) and scientific applications (statistical physics, biology). Although graphs are typically depicted by nice drawings, it quickly becomes a jumble of wires when their size increases. Helping to understand the information content of such graphs is an important challenge that cannot be addressed purely by smart visualization methods using different layouts.
We will focus on graphs issued from observation data in experimental sciences. In this domain, graphs (also called networks) are a way to integrate various sources of data and the role of compression is to help scientists facing this type of data to extract from them new knowledge. This general goal can take a number of forms: to simplify visualization, to show the global graph structure, to find general association rules, or simply to look for better efficiency for other analysis tasks (e.g. pattern matching).

Objective: The thesis will study ways to summarize a graph that preserves its path structure in the context of experimental sciences, and more particularly in biology. At a high level, the central issue of graph compression may be viewed as a clustering issue. Preserving the global structure of the graph may be achieved by keeping the graph representation and replacing nodes by more abstract nodes corresponding to node subsets. Two abstract nodes are linked by an abstract edge if all their elements are linked by an edge. This corresponds to a subgraph called biclique. Clusters of nodes are typically nodes that share common edges and common properties [1].
Lossless graph compression assumes that all edges are preserved and this raises the problem of covering a graph with a subset of its bicliques. The simplest idea is to look for a node partition, i.e. a set of mutually disjoint edge subsets. This condition is too strict in practice and people generally look for biclique partitions, i.e. sets of bicliques that are mutually edge-disjoint. This problem is known to be NP-complete, even if restricted to bipartite graphs, but fixed-parameter tractable [2]. The paper [3] proposes a bioinformatics tool (Power Graph Analysis) which outputs a biclique partition of a graph such that node subsets form a hierarchy, using a greedy algorithm. It has been applied with success on protein interaction networks [4], drug-target-disease network [5] and gene regulatory networks [6]. It will be the starting point of the thesis. A number of questions have still to be addressed, for instance:

- Among all bicliques, maximal ones (also called concepts) form an interesting subset that is the subject of formal concept analysis (FCA). A nice theoretical framework exists for describing them within a lattice structure [7]. How to generate the graph cover using only concepts and a simple algebra for combining them ? Is it possible to solve this covering problem as a search in the concept lattice ?
- Most of existing methods use a greedy search algorithm that is looking for a local optimum in terms of a minimum number of bicliques used. Covering is a purely combinatorial problem for

which constraint-based methods seem more adapted since it allows a search in a larger hypothesis space and lead to better solutions. We propose to express this problem in the framework of ASP (Answer Set Programming) ASP offers a high level language that represents a problem specification as a set of simple logic constraints[8]. Once the problem has been correctly represented, a solver is in charge of the resolution so that it is not necessary to describe how to solve the problem.

- Instead of looking for node similarity, it is possible to search for subgraph similarity. Each abstract node becomes a subgraph that can itself be compressed. The search of interesting subgraphs is the subject of graph mining techniques [9] and it will provide a basis to achieve better compression rate on biological graphs.

Environment: The thesis is proposed in a bioinformatics team, Dyliss, which is concerned by modeling issues of complex molecular mechanisms in living cells using qualitative formalizations. Through multiple collaborations with biological labs, it has access to various networks (protein-protein interaction network, regulation networks, signaling networks…) that will be used to test concepts and tools developed during the thesis. There exists a resource center, Genouest, giving access to biological databases and bioinformatics software, and offering cluster and cloud computing facilities. The work includes a collaboration with a German Computer Science team in Potsdam that has designed one of the best ASP solver to date.

Key words: Graph, Compression, NP-complete problems, Combinatorial search, Discrete optimization, Formal concept analysis, Answer set programming, Bioinformatics

First references:

[1] Zhang Y, Phillips CA, Rogers GL, Baker EJ, Chesler EJ, Langston MA. 2014. On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinformatics. 15:110. doi: 10.1186/1471-2105-15-110.
[2] Fleischner H, Mujuni E, Paulusma D, Szeider S. 2009. Covering graphs with few complete bipartite subgraphs. Theoretical Computer Science, Vol. 410 n°21-23, pp 2045-2053,http://dx.doi.org/10.1016/j.tcs.2008.12.059.
[3] Royer L, Reimann M, Andreopoulos B, Schroeder M (2008) Unraveling Protein Networks with Power Graph Analysis. PLoS Comput Biol 4(7): e1000108. doi:10.1371/journal.pcbi.1000108
[4] Royer L, Reimann M, Stewart AF, Schroeder M. Network compression as a quality measure for protein interaction networks. PLoS One. 2012;7:e35729.
[5] Daminelli S1, Haupt VJ, Reimann M, Schroeder Drug repositioning through incomplete bi-cliques in an integrated drug-target-disease network. Integr Biol (Camb). 2012 Jul;4(7):778-88. doi: 10.1039/c2ib00154c
[6] Taylor-Teeples M, Lin L, de Lucas M, Turco G et al., An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature. 2015 Jan 29;517(7536):571-5.
[7] M.Ganter B, Wille R. 1997. Applied lattice theory: Formal concept analysis. In General Lattice Theory, G. Grätzer editor, Birkhäuser. http://www.math.tu-dresden.de/~ganter/psfiles/concept.ps
[8] Lifschitz V. 2008. What is answer set programming?. In *Proceedings of the 23rd national conference on Artificial intelligence* (AAAI'08), Anthony Cohn (Ed.), Vol. 3. AAAI Press 1594-1597.
[9] S. Parthasarathy, S. Tatikonda, D. Ucar A Survey of Graph Mining Techniques for Biological Datasets. Managing and Mining Graph Data Advances in Database Systems Vol 40, 2010, pp 547-580

Links Dyliss: http://www.irisa.fr/dyliss Answer Set Programming: http://potassco.sourceforge.net/