INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team Gyroweb

Dynamic graphs and large networks

Rocquencourt

THEME 1B

Table of contents

1. Team

Gyroweb is a joined team between INRIA, CNRS and Paris 7 University, through the "laboratoire d'informatique algorithmique, fondements et applications", LIAFA *(UMR 7089).*

Head of project-team
Laurent Viennot [Research scientist (partner)]

Vice-head of project team
Dominique Fortin [Research scientist (partner)]

Administrative assistant
Danielle Croisy [shared time (with Hipercom)]

Staff member Paris 7 University
Yacine Boufkhad [Assistant professor]

Staff member CNRS
Matthieu Latapy [Research scientist (partner)]

Ph. D. student
Anh-Tuan Gai [BDI]
Jean-Loup Guillaume [ENS Cachan]
Fabien Mathieu [ENS Ulm]

Student intern
Stevens Le-Blond [Epitech]

2. Overall Objectives

The main objective of the project is to study the structure and the dynamics of the graphs appearing in large networks. The web graph defined by the web pages and the hyper-links between them is one of our main application field. Other natural subjects of interest include the Internet graph (defined by the connections between Internet routers), peer-to-peer networks (where a logical network links the peers together) and social networks.

The first problem comes from measuring such large structures. Usually made through crawling, the process of discovering the large graphs listed above is long and incomplete. How to define precisely what is measured, quantifying how accurate is a measure are still open problems.

The natural follow-up is to model such large networks. Recent work has exhibited power laws in degree distributions of these graphs. However, few efforts have been made on modeling the dynamics of these evolving graphs.

Finally, these graphs are linked to algorithms and protocols as page-rank for the web graph, routing for the Internet graph and indexing for peer-to-peer networks. Optimizing them by using the knowledge of the structure of the underlying graph is one of our goals.

3. Scientific Foundations

Key words: *graph algorithms*, *network protocols*, *stochastic models*, *matrix analysis*, *dynamic graphs*.

The main competencies of the team are graph algorithms, networking protocols, matrix analysis, and stochastic models.

3.1.1. Networking protocols and graph theory

The design of networking protocols (such as ad hoc routing protocols or file sharing peer-to-peer protocols) often rely on distributed graph algorithms. Many invariants for the good functioning of them can be expressed as graph theoretical properties. Moreover, many observations of large graphs rely on the underlying protocols

allowing to measure them (such as HTTP for crawling the web graph, or BGP and ICMP for discovering the Internet graph). Knowledge of these protocols and network constraints are fundamental when modeling these graphs.

3.1.2. Stochastic graph models

The popular random graph model of Erdõs and Rènyi does not capture the properties observed in real world complex networks (web graph, social graphs, etc). These properties are in general a power law degree distribution, a low diameter, and a high clustering. Consequently, there is a need to find new models for these graphs. The aim at finding such models is to understand the basic mechanisms behind their particular structure to predict their future evolution and to randomly generate graphs having the same properties. Although some results on graphs with power law degree distribution have been published, much still has to be done in this area.

3.1.3. Matrix and spectral analysis of large graphs

Matrix analysis reveals some important properties of large graphs through their adjacency matrix. The pagerank for example can be viewed as an eigenvector of a normalization of the adjacency matrix of the web graph. More generally, structural analysis splits into spectral analysis and actual domain approximation. In the former, we relate eigenvalues and eigenvectors to properties like maximum clique, maximum cut, etc..., while in the latter we devise simpler structures that leads to close approximations of general ones. Due to large sizes, both approaches are rather complementary than competitors.

3.1.4. Dynamic graphs

An abundant literature around fully dynamic algorithms treats the problem of updating a graph computation after slight modification of the graph. Another approach resides in considering a dynamic graph as an evolving structure and trying to characterize the properties of this evolution.

4. Application Domains

Key words: *mapping the web*, *Internet reliability*, *peer-to-peer protocols*.

Application domains include mapping the web, evaluating Internet reliability, and the design of efficient peer-to-peer protocols.

- The main application of studying the web graph resides in evaluating the importance of web pages (as popularized by the PageRank of Google). A long term application goal is to map the web: that is to identify automatically sites and links between them.
- Modeling the Internet graph main application is to allow realistic simulation of Internet protocols. Another interesting field resides in evaluating the reliability of the Internet connectivity.
- Peer-to-peer protocols are based on a all equal paradigm that allows to design highly reliable and scalable applications. Besides the file sharing application, peer-to-peer solutions could take over in web content dissemination resistant to high demand bursts or in mobility management.

5. Software

The team develops internal tools for crawling and generating large random networks.

- Jean-Loup Guillaume has developed random graph generators http://www.liafa.jussieu.fr/~guillaume/programs/.
- The Web graph encodings we propose in [2] are used by other researchers in various softwares.
- The team has developed tools for crawling the web. Libraries for the free crawler “larbin” of Sébastien Ailleret are available at http://menetou.inria.fr/soleil/database.html.
- Stevens Le-Blond is developing a crawler for the Gnutella network.

6. New Results

6.1. Realistic Modeling of Complex Networks

Participants: Jean-Loup Guillaume, Matthieu Latapy.

Key words: *random graphs*, *bipartite networks*, *cliques*.

It appeared recently that the classical random graph model used to represent real-world complex networks does not capture their main properties. Since then, various attempts have been made to provide accurate models. We present surveys of these works and related questions in [1][10][9].

We present in [19] a way to see any complex network as a bipartite graph, and we propose in [18] the first model which achieves the following challenges: it produces graphs which have the three main wanted properties (clustering, degree distribution, average distance), it is based on some real-world observations, and it is sufficiently simple to make it possible to prove its main properties. This model consists in sampling a random bipartite graph with prescribed degree distribution. Indeed, we show that any complex network can be viewed as a bipartite graph with some specific characteristics, and that its main properties can be viewed as consequences of this underlying structure. We also propose a growing model based on this observation, which fits surprisingly well the properties met in practice.

6.2. Web Graph

Participants: Yacine Boufkhad, Fabien Mathieu, Laurent Viennot, Mohamed Bouklit [LIRMM, Montpellier].

Key words: *web graph*, *PageRank*, *observable web*.

Many results of the team concern the study of the web graph.

6.2.1. The Observable Web

In [13], we estimate how accurate is the view of the web obtained by crawling. Our approach is to compare crawling to other ways of discovering the web (mainly server or proxy logs of web surfers activity). This work is a first step towards identifying the observable web. The latter being defined as the part of the web that can be discovered automatically.

6.2.2. Local Structure in the Web

The web graph has been widely adopted as the core describing the web structure. However, little attention has been paid to the relationship between the web graph and the location of the pages. Locality in the web can be further modeled by the clustered graph induced by the prefix tree of URLs [12]. The web tree's internal nodes are the common prefixes of URLs and its leaves are the URLs themselves. A prefix ordering of URLs according to this tree allows to observe local structure in the web directly on the adjacency matrix M of the web graph. M splits in two terms : $M = D + S$, where D is diagonal by blocks and S is a very sparse matrix. The blocks of D that can be observed along the diagonal are sets of pages strongly related together.

6.2.3. Ranking of Web Pages

The PageRank is a distribution of probability on the web pages computed from the web graph. It allows efficient estimation of pages importance. In [11], we show how the PageRank can split into two terms, an internal and an external PageRank. These two PageRanks allow a better understanding of the PageRank signification inside and outside a site. A first application of this model is a local algorithm for estimating the global PageRank inside a site with local knowledge.

The PageRank is better modeled by a random surfer Markov process. A model including a back step (motivated by the back button of web browsers) is proposed in [8].

6.3. Graph and Matrix Algorithms

Participants: Anh-Tuan Gai, Dominique Fortin, Laurent Viennot.

Key words: *maximum clique*, *common connected components*, *ad hoc routing*.

6.3.1. Maximum Clique

Participants: Dominique Fortin, Ider Tseveendorj [Laboratoire de Mathématiques d'Orléans].

In [17] we carry on spectral analysis of maximum clique problem and introduce Perron's regularization as a non diagonal perturbation; it puts further shed on resolvent sensitivity analysis and eigenvector space.

6.3.2. Graphs Comparisons

Participants: Anh-Tuan Gai, Michel Habib [LIRMM, Montpellier], Christophe Paul [LIRMM, Montpellier], Mathieu Raffinot [Laboratoire Génome et Informatique de Tour].

In [15], we study the first problem of graphs comparison: The Common Connected Problem (CCP). CCP consists in identifying common connected components of two or more graphs on or reduced to the same vertex set. More formally, let $G_1(V, E_1)$ and $G_2(V, E_2)$ be two such graphs and let $G_1[V']$ and $G_2[V']$ be two subgraphs induced by a set of vertices $V' \in V$. If $G_1[V']$ and $G_2[V']$ are both connected, V' is called a common connected component. CCP is the identification of such maximal components (considering the inclusion order), that form a partition of V. Let $n = |V|$ and $m = |E_1| + |E_2|$. We present the first, to our knowledge, non-trivial algorithm solving CCP, running in $O(n \log n + m \log^2 n)$ worst case time.

6.3.3. Ad hoc routing

Participants: Laurent Viennot, Géraud Allard [Hipercom, Inria], Thomas Clausen [Hipercom, Inria], Philippe Jacquet [Hipercom, Inria].

In [6], we show how to optimize reactive ad hoc routing protocols through multipoint relaying, a proactive flooding technique. In [4], we compare control traffic overhead of main ad hoc routing protocol families with regard to mobility and data traffic activity.

6.4. Discrete metrics

Participants: Dominique Fortin, Pascal Préa [Laboratoire d'Informatique Fondamentale de Marseille].

Key words: *Robinson*, *Supnick*, *Kalmanson*, *seriation*.

In clustering, an important issue deals with approximating a given dissimilarity with a special structured dissimilarity; in [20][16], we extend the well known Robinson property defined on 3 points to 4 points among a linear extension of all points. It yields a complete characterization of such bottleneck matrices; moreover, OGFs are given to estimate the rate of growth in terms of the size and their subclass. Despite this complete characterization, extremal rays of the corresponding cones remain implicitly described; therefore, future work should be undertaken to answer the question whether optimization on these cones has a chance to be polynomial. Given an $n \times n$ dissimilarity D, it is Kalmanson whenever $c_{ik} + c_{jl} \geq c_{ij} + c_{kl}$ and $c_{ik} + c_{jl} \geq c_{il} + c_{jk}$ for all $1 \leq i < j < k < l \leq n$; it is Supnick whenever $c_{ij} + c_{kl} \leq c_{ik} + c_{jl} \leq c_{il} + c_{jk}$ for all $1 \leq i < j < k < l \leq n$. These cones follow the following inclusion map: Robinson's cone $\subset$ Supnick's cone $\subset$ Kalmanson's cone.

6.5. SAT related problems

Participants: Yacine Boufkhad, Olivier Bailleux [LERSIA, Bourgogne], Olivier Dubois [LIP6], Jacques Mandler [LIP6].

Key words: *satisfiability*, *phase transition*, *upper bound*, *encoding*, *cardinality constraints*.

6.5.1. The Satisfiability phase transition

In [14], we present a new structural (or syntactic) approach for estimating the satisfiability threshold of random 3-SAT formulae. We show its efficiency in obtaining the best upper bound for 3-SAT phase transition by a jump from the previous upper bounds, lowering them to 4.506. The method combines well with other techniques, and also applies to other problems, such as the 3-colorability of random graphs.

6.5.2. Encoding problems into satisfiability

In [7], we address the encoding into CNF clauses of Boolean cardinality constraints that arise in many practical applications. The proposed encoding is efficient with respect to unit propagation, which is implemented in almost all complete CNF satisfiability solvers. We prove the practical efficiency of this encoding on some problems arising in discrete tomography that involve many cardinality constraints.

In [3], we address the problem of reconstructing a bidimensional pattern given its approximate horizontal and vertical projections. We deal with the case of patterns which have to be horizontally and vertically convex and the case of patterns which have to be moreover connected, so-called convex polyominoes. We show that in both cases, the problem of reconstructing a pattern can be transformed into a Satisfiability Problem. This is done in order to take advantage of the recent advances in the design of solvers for the Satisfiability problem. We show, experimentally, that by adding two important features to CSAT (an efficient SAT solver), optimal patterns can be found efficiently if there exist feasible ones.

8. Other Grants and Activities

8.1. National initiatives

8.1.1. Dynamo CNRS Investigation Grant

The Dynamo CNRS Investigation Grant leaded by Pierre Fraigniaud (LRI) focuses on structural and algorithmic aspects of dynamic networks. Matthieu Latapy is the head of the web graph theme.

8.1.2. Gap Ministry Grant

The team collaborates to the GAP (Graphs, Algorithms and Probability) national project.

8.1.3. PairAPair Ministry Grant

Laurent Viennot is the head of the PairAPair national project. Anh-Tuan Gai is funded by this project.

Peer-to-peer networks have become the heaviest source of traffic in the Internet through the use of file sharing applications (such as Gnutella or Kazaa for example). However the protocols behind these applications are still too greedy and waste a lot of the Internet resources. On the other hand, theoretical solutions based on distributed hash tables exist but cannot be used practically. The PairAPair project aims at bridging efficient theoretical solutions to practical applications such as file sharing.

The main goal of the project concerns the conception of peer-to-peer protocols. A first approach consists in optimizing algorithms for existing protocols (without changing the communication rules). Another way consists in developing new protocols based on efficient theoretical solutions. Also important aspects of peer-to-peer networks concerns ethics: how to accept sharing one's resources if they can be used for non moral purposes? Designing protocols allowing the respect of certain rules will be another goal of the PairAPair project. Finally, analyzing and optimizing protocols requires models. For that purpose, crawling of existing peer-to-peer networks is envisioned.

The PairAPair project gathers members of four teams: Gyroweb (INRIA–LIAFA), GraFComm (LRI), Graphes (LABRI) and Hipercom (INRIA). More information is vailable at http://gyroweb.inria.fr/pairapair/.

8.2. European initiatives

8.2.1. IO Network of Excellence

Dominique Fortin was one the leaders of the IO proposition of European *Network of Excellence (NoE)*.

For FP6, he has spent much time on a proposal for infrastructure optimization with the project coordinator (JF Maurras from the Laboratoire d'Informatique Fondamentale de Marseille) to structure ninety researchers from the world (not only EC); despite this big effort, our proposal was rejected. Here is a short description of our proposal: there is an increasing demand to cope with the complexity of real-world applications. Nowadays computer platforms and their usage have emerged into exponentially growing interconnected components. The

vast amount of data that is produced by the world wide web for mobile phones to name the most hyperactive, leads to tremendous infrastructure optimization problems where optimizing the cost involved could lead to valuable savings. In such a context, it is necessary to study and propose exact methods and heuristics with provable performance guarantees. While, the nature of these problems have long been known, the sizes involved and some challenging effects need to settle cooperating forces to face them. A Network of Excellence should result in an opened working platform to concentrate knowledge of major research questions (to balance the provision made by the web to disseminate information in a definitely unstructured fashion) among a few reference sites only. It should also provide researchers with an European framework between universities to carry out joint works within NoE partners.

8.2.2. Objectives

Proposal objectives range from general to operational levels as

- elaborating the optimization models for difficult problems occurring in: communication, transportation, networking, data analysis...,
- classifying addressed problems: network and optimization design, combinatorial and polyhedral optimization, scheduling, facility location, large-scale optimization... (to concretely exemplify this item, say among others, graph partitioning, one edge network failure, scheduling, 2-edge connectivity, network design and network optimization),
- developing methods for solving such problems: designing exact and approximation algorithms,
- developing integer programming and non linear numerical methods and integrating different methods into decision support systems,
- developing tools for different optimization methods (e.g. libraries for Branch and Bound/ Cut/Price, algorithmics for combinatorial optimization...)

As a long term structuring effect, we build our NoE on top of a proposal lead common to all teams, namely the Peer-to-Peer Computing paradigm which is in the heart of Gyroweb application field.

8.2.3. SAGA Network of Excellence

The team has also participated in the SAGA proposition of European *Network of Excellence* that was centered on algorithmic aspects of networking.

9. Dissemination

9.1. Conferences, meetings and tutorial organization

9.1.1. Seminar

Matthieu Latapy organizes with Clémence Magnien from LIX (École Polytechnique) a monthly afternoon seminar (three speakers) on *Graphs, Networks and Modeling*. These meetings have a regional scope.

9.1.2. Graphs, Networks and Modeling Meeting

A two day meeting is also organized in conjunction with this seminar. It will take place at ESPCI in Paris on December the 17th and the 18th. 33 abstracts were reviewed and 70 participants are registered.

9.2. Teaching

D.E.A. Algorithmics, Paris 7, 6 and 11 Universities, Ecole Polytechnique, ENS Ulm, ENS Cachan, ENST Laurent Viennot is teaching network algorithmics (14 hours).

D.E.S.S. Fundamental Software, Paris 7 University Matthieu Latapy has given a professional conference (2 hours).

ENS Cachan Matthieu Latapy is teaching on large graphs in practice (7 hours).

Ecole Polytechnique Laurent Viennot is teaching foundations of computer science, java and networks (90 hours). Anh-Tuan Gai is teaching foundations of computer science (30 hours).

Master Epita Matthieu Latapy is teaching C and imperative programming (16 hours).

Maîtrise Marseille University Dominique Fortin is teaching software engineering (30 hours).

D.U.T. Paris 7 University Yacine Boufkhad is teaching scientific computer science and networks (192 hours). Anh-Tuan Gai is teaching C++ (30 hours).

D.E.U.G. Paris 7 University Jean-Loup Guillaume is teaching unix, automata and programming (64 hours).

D.E.U.G. Montpellier 2 University Fabien Mathieu is teaching scheme in DEUG and systems in Licence (64 hours).

10. Bibliography

Major publications by the team in recent years

[1] J.-L. GUILLAUME, M. LATAPY. *The Web Graph: an Overview.* in « 5es rencontres francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL'2002) », 2002.

[2] J.-L. GUILLAUME, M. LATAPY, L. VIENNOT. *Efficient and Simple Encodings for the Web Graph.* in « The Third International Conference on Web-Age Information Management (WAIM) », august, 2002, Beijing.

Articles in referred journals and book chapters

[3] Y. BOUFKHAD, O. DUBOIS, M. NIVAT. *Reconstructing (h, v)-convex 2-dimensional patterns of object s from approximate horizontal and vertical projections.* in « Theoretical Computer Science », volume 290(3), 2003, pages 1647-1664.

[4] T. CLAUSEN, P. JACQUET, L. VIENNOT. *Analyzing Control Traffic Overhead versus Mobility and Data Traffic Activity in Mobile Ad-hoc Network Protocols.* in « ACM Wireless Networks journal (Winet) », number 4, volume 10, July, 2004, to appear.

[5] D. FORTIN, I. TSEVEENDORJ. *Global Optimization and Multiknapsack: a percolation algorithm.* in « European Journal of Operational Research », 2003, available online 11 March 2003.

Publications in Conferences and Workshops

[6] G. ALLARD, P. JACQUET, L. VIENNOT. *Ad hoc routing protocols with multipoint relaying.* in « 5es rencontres francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL'2003) », 2003.

[7] O. BAILLEUX, Y. BOUFKHAD. *Efficient CNF encoding of boolean cardinality constraints.* in « Ninth International Conference on Principles and Practice of Constraint Programming », 2003.

[8] M. BOUKLIT, F. MATHIEU. *Effet de la touche Back dans un modèle de surfeur aléatoire : application à PageRank.* in « Journées francophones de la toile (JFT'2003) », 2003.

[9] J.-L. GUILLAUME, M. LATAPY. *Modèles pour les topologies réalistes.* in « 5es rencontres francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL'2003) », 2003.

[10] J.-L. GUILLAUME, M. LATAPY. *Topologies d'Internet et du Web : mesure et modélisation.* in « Colloque *Mesure de l'Internet* (CMI'2003) », 2003.

[11] F. MATHIEU, L. VIENNOT. *Aspects locaux de l'importance globale des pages web.* in « 5es rencontres francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL'2003) », 2003.

[12] F. MATHIEU, L. VIENNOT. *Local Structure in the Web.* in « 12-th international conference on the World Wide Web », 2003, http://www2003.org/cdrom/papers/poster/p102/p102-mathieu.htm, poster.

Internal Reports

[13] Y. BOUFKHAD, L. VIENNOT. *The Observable Web.* Technical report, number RR-4790, INRIA, Rocquencourt, April, 2003, http://www.inria.fr/rrrt/rr-4790.html.

[14] O. DUBOIS, Y. BOUFKHAD, J. MANDLER. *Typical random 3-SAT formulae and the satisfiability threshol d.* Technical report, ECCC, 2003.

[15] A. GAI, M. HABIB, C. PAUL, M. RAFFINOT. *Identifying Common Connected Components of Graphs.* Technical report, number RR-LIRMM 03-016, LIRMM, 2003, http://www.lirmm.fr/~paul/Biblio/Postscript/RR-LIRMM03016.ps.

Miscellaneous

[16] D. FORTIN, P. PRÉA. *The Cone of Kalmanson dissimilarities: extremal Rays and Generalizations.* 2003, submitted.

[17] D. FORTIN, I. TSEVEENDORJ. *Maximum clique and conic programming: from diagonal to offdiagonal regularizations.* 2003, submitted.

[18] J.-L. GUILLAUME, M. LATAPY. *Bipartite Graphs as Models of Complex Networks.* 2003, submitted.

[19] J.-L. GUILLAUME, M. LATAPY. *Bipartite structure of all Complex Networks.* 2003, submitted.

[20] P. PRÉA, D. FORTIN. *The Cone of Supnick Bottleneck matrices.* 2003, submitted.