



INSTITUT DE FRANCE
Académie des sciences

Avancées en Sciences de l'Information présentées par leurs auteurs



© Institut de France/FESSY G.

**Séance publique de
l'Académie des sciences**

Mardi 9 octobre 2007 à 14h30

**Grande salle des séances
23 Quai de Conti – 75006 Paris**

Séance publique consacrée à six avancées en sciences de l'information
présentées par leurs auteurs*

Programme

- 14h30 **Introduction par François Baccelli**, de l'Académie des sciences.
- 14h35 **Georges Gonthier**, Centre de recherche commun INRIA-Microsoft Research
"Le théorème des quatre couleurs: ingénierie d'une preuve formelle".
- 15h05 **Gilles Schaeffer**, CNRS - Laboratoire d'informatique de l'Ecole polytechnique
"Arbres couvrants canoniques, des géométries aléatoires à la compression de maillages".
- 15h35 **Patrice Abry**, CNRS - ENS Lyon
"Invariance d'échelle pour la modélisation du trafic Internet".
- 16h05 **Laurent Massoulié**, Thomson
"Diffusion épidémique d'information dans les réseaux pair-à-pair".
- 16h35 **Frédéric Cazals**, INRIA Sophia, projet Geometrica
"Modèles et algorithmes pour la description des interactions macro-moléculaires : le triptyque biophysique - géométrie - statistiques".
- 17h05 **Sylvain Soliman**, INRIA Rocquencourt, projet Contraintes
"Langages formels pour la biologie systémique dans la machine abstraite biochimique BIOCHAM".

*** Entrée libre**

Informations : Service des Colloques de l'Académie des sciences – 01 44 41 43 82 / 43 87 / 44 61
fabienne.bonfils@academie-sciences.fr - <http://www.academie-sciences.fr>

Le théorème des quatre couleurs: ingénierie d'une preuve formelle

Georges GONTHIER et Benjamin WERNER

Centre de Recherche Commun INRIA - Microsoft Research

La résolution du problème séculaire des quatre couleurs par Appel et Haken en 1976, à l'aide d'un calcul compliqué sur ordinateur, fut le début d'une polémique sur l'utilisation des ordinateurs en mathématiques : un tel calcul pouvait-il vraiment avoir valeur de preuve ? Trente ans plus tard, nous pouvons enfin répondre par l'affirmative: il est possible de construire effectivement une preuve complètement formelle du théorème des quatre couleurs, dont la correction ne repose que sur des calculs simples et reproductibles, et dont la composition apporte un nouvel éclairage sur le problème.

Pour valider une preuve sur ordinateur il faut démontrer que la machine exécute exactement les calculs exigés par la démonstration. Pour cela il suffit en principe de décrire aussi bien les calculs que la démonstration avec une rigueur formelle complète ; mais d'aucuns considéraient que ceci était inutile et infaisable en pratique. Nous renversons cette thèse en exhibant une preuve formelle du théorème des quatre couleurs, rigoureusement et efficacement vérifiable, qui éclaire et approfondit la preuve classique dont elle s'inspire (due à Robertson, Sanders, Seymour et Thomas). En outre, comme la preuve est entièrement vérifiée par le calcul, il n'est pas nécessaire de la relire en entier pour s'assurer de l'authenticité du résultat - il suffit d'en lire les 0,2% qui correspondent à l'énoncé mathématique du théorème.

Nos travaux s'appuient sur un ensemble de résultats en logique, informatique théorique, algorithmique, et génie logiciel. En particulier, nous avons construit notre preuve dans le système Coq développé à l'Inria. Le système Coq est un **assistant à la preuve** – un programme qui permet de transformer automatiquement un schéma de preuve en une preuve formelle détaillée et rigoureusement vérifiée. La logique de Coq permet d'intégrer efficacement l'exécution de programmes informatiques à des énoncés mathématiques. Ceci nous a permis non seulement de démontrer la correction de calculs complexes, comme la réductibilité et l'inévitabilité dans la preuve de Appel et Haken, mais aussi de préciser et de simplifier des énoncés purement mathématiques, notamment en théorie des graphes.

Nous avons aussi utilisé des techniques issues du génie logiciel pour maîtriser la complexité de la formalisation. Par exemple, nous avons cloisonné la preuve, isolant les parties algorithmiques, combinatoires, et topologiques. Dans la partie algorithmique nous avons remplacé les méthodes ad hoc par des algorithmes connus, issus pour la plupart de travaux sur la vérification formelle. Dans la partie combinatoire nous avons substitué des hypercartes aux graphes, ce qui nous a permis de remplacer des appels un peu flous à l'intuition géométrique par des identités algébriques rigoureuses. Dans la partie topologie, nous avons découvert que la preuve du théorème des quatre couleurs ne passait pas par la construction d'un graphe dual, et donc ne dépendait pas du théorème de Jordan.

Nous pensons que nos travaux préfigurent plusieurs évolutions, à la fois en mathématiques et en informatique. La faisabilité de preuves formelles irréfutables va en faire un moyen effectif de résolution de controverses mathématiques, dont le théorème de Hales (conjecture de Kepler) sera sans doute le prochain exemple. L'intérêt intrinsèque de ces preuves va en faire un nouvel outil pour l'investigation des problèmes mathématiques, qui amènera sans doute de nouvelles découvertes. Enfin, nous avons montré qu'il était possible et même fructueux d'aborder des preuves complexes par le génie logiciel, ce qui nous laisse penser que les démonstrations mathématiques pourraient jouer un rôle plus important dans le développement de logiciel.

Arbres couvrants canoniques, des géométries aléatoires à la compression de maillages

Gilles SCHAEFFER

Laboratoire d'informatique de l'Ecole Polytechnique

Les parcours de graphes en largeur ou en profondeur sont parmi les outils les plus fondamentaux de l'algorithmique. J'en exposerai une propriété élégante qui a permis à la fois d'avancer dans la compréhension de certaines géométries aléatoires considérées en physique statistique et de proposer des codages optimaux, au sens de la théorie de l'information, pour les structures des maillages de surfaces utilisées classiquement en infographie.

L'idée est d'utiliser ces parcours pour mettre à jour des arbres couvrants canoniques remarquables dans les triangulations planes, ou plus généralement dans les cartes sous-jacentes aux maillages de surfaces. La notion de *carte* est pour les objets géométriques bidimensionnels (les surfaces discrétisées, les pavages, etc.) ce que la notion combinatoire d'*arbre* est aux structures arborescentes: une abstraction mathématique extrêmement commode pour en étudier les propriétés génériques. Le fait d'associer un arbre couvrant à une carte pour en coder la structure est classique. L'originalité des arbres obtenus par mes algorithmes est qu'on puisse décrire facilement l'ensemble des arbres utiles (ainsi les triangulations sont associés aux arbres ordonnés dont chaque sommet est voisin d'exactly deux feuilles, facilement codés par une grammaire algébrique).

Je présenterai deux applications de ces idées. La première concerne la compression de maillages. La compression d'un maillage repose classiquement sur un codage de la carte sous-jacente par un mot binaire, associé à une liste de coordonnées géométriques. Le codage de la carte a fait l'objet d'une série d'articles améliorant les taux de compression et nous avons montré avec D. Poulalhon comment les arbres canoniques permettent, en codant une triangulation par son arbre et l'arbre par un mot binaire, d'atteindre asymptotiquement le taux de compression optimal (au sens de la théorie de l'information: aucun codeur ne peut faire significativement mieux en toute circonstance).

Une seconde application que j'illustrerai est l'étude de la géométrie intrinsèque des cartes aléatoires. La distribution uniforme sur les cartes planaires de taille fixée est un modèle de surfaces aléatoires naturel combinatoirement et étudié en physique statistique en lien avec la discrétisation de la gravité quantique. Le fait que mes arbres couvrants conservent l'information des distances entre sommets de la carte qu'ils codent nous a permis, avec P. Chassaing de démontrer le résultat, prédit par les physiciens, que ces distances se comportent en $n^{1/4}$ dans une carte de taille n . Nous avons ainsi mis à jour un lien profond entre ces géométries aléatoires et des superprocessus étudiés en probabilité, ce qui a ouvert la voie à la construction de limites continues des cartes aléatoires (cf. notamment les articles récents de J.F. Le Gall).

Articles exposés:

- Dominique Poulalhon et Gilles Schaeffer, Optimal Coding and Sampling of Triangulations, *Algorithmica*, **46**, 505-527 (2006).
- Philippe Chassaing et Gilles Schaeffer, Random planar lattices and integrated superBrownian excursion, *Probab. Theory Relat. Fields*, **128**, 161--212 (2004).

Invariance d'échelle pour la modélisation du trafic Internet

Patrice ABRY

Equipe Signaux, Systèmes et Physique
CNRS, Laboratoire de Physique, Ecole Normale Supérieure de Lyon

Internet devient l'ultime medium de communication, universel, rassemblant des échanges d'informations de quasi toutes natures, textuelle, orale (audio), image (vidéo)... sous quasi toutes les formes envisageables (en direct, en différé, par interrogations de sites et bases de données référencés, par échanges entre internautes (pairs à pairs), par inter-activités. De plus, les échanges s'effectuent dans les mêmes conditions (simplicité, efficacité,...) et aussi rapidement entre deux internautes, quels que soient leurs proximité ou éloignement géographiques réels, abolissant ainsi, ou renouvelant profondément, les conceptions classiques de temps et d'espace.

Le développement sans cesse plus rapide, plus radical, du réseau Internet (que ce soit dans ces aspects matériels ou dans celui des applications et usages qu'il permet) fait l'objet d'enjeux scientifiques, technologiques, économiques et sociaux majeurs. Internet, fascinant parce que grand système complexe, fortement hétérogène, en perpétuels évolutions et changements, donc vivant et dynamique est actuellement élevé au statut d'objet d'études et de recherche par de multiples communautés scientifiques - Informatique, Réseaux et Communications, Traitement du Signal, Mathématiques, Economie, Sociologie...

La présentation proposée ici s'intéresse spécifiquement à la caractérisation statistique et à la modélisation stochastique des séries temporelles matérialisant les flux d'informations échangés sur Internet. Les motivations de telles études résident dans l'usage qui peut être fait a posteriori de ces modélisations pour assurer le bon fonctionnement du réseau (éviter les pertes, minimiser les délais de transmission), pour penser et optimiser son développement (topologie, dimensionnement...), pour en assurer la sécurité (détection d'anomalies, de pannes). Au-delà de ces objectifs appliqués clairement identifiés, l'étude du trafic Internet constitue également un défi scientifique motivant ; hétérogénéité et complexité induisent des propriétés statistiques non standards et difficiles (processus non stationnaires, fortement non gaussiens éventuellement à *ails lourdes*, à longue mémoire, invariant d'échelle...) qu'il faut savoir analyser et modéliser.

Nous montrerons notamment comment les notions d'invariance d'échelle, d'auto-similarité et de longue mémoire, constituent des propriétés fortement structurantes de la description du trafic Internet. Nous illustrerons également les difficultés qu'elles posent à l'analyse. Les modèles fractals et multifractals, en plein essor théorique et très fréquemment avancés dans de nombreux domaines scientifiques, ont été mis en avant pour décrire le trafic Internet. Nous montrerons qu'une analyse expérimentale fine du trafic Internet ne valide pas ce modèle. En place, nous proposons d'utiliser un modèle de processus ponctuel, dit *de grappe*, qui rend compte, avec pertinence de l'invariance d'échelle, observée dans les données, tout en renouant avec le caractère naturellement ponctuel du trafic Internet, consistant *in fine* en la circulation d'une collection de paquets IP (*Internet protocol*), organisés en connexions.

Diffusion épidémique d'information dans les réseaux pair-à-pair

Laurent MASSOULIE

Laboratoire de recherche Parisien, Thomson

Le pair-à-pair aujourd'hui : Plus de 80% du trafic écoulé sur l'Internet provient de systèmes pair-à-pair de partages de fichiers. Au delà du partage de fichiers, les dernières générations de systèmes pair-à-pair permettent de visualiser un flux vidéo en temps réel sur ordinateur.

Dans les prochaines années, les systèmes pair-à-pair déployés sur l'Internet peuvent devenir le principal canal de diffusion de données multimédia.

Les enjeux pour la recherche : L'analyse des possibilités et des limites fondamentales des capacités de diffusion de tels systèmes constitue donc un enjeu industriel majeur.

Il s'agit d'identifier des algorithmes distribués permettant une diffusion avec la meilleure performance possible. Seules des solutions distribuées sont faisables: les systèmes pair-à-pair mettent en jeu des milliers de composants (les utilisateurs), qui rejoignent ou quittent le système à leur gré. Un problème corollaire est le suivant: une solution distribuée est-elle intrinsèquement moins performante qu'une solution centralisée?

Principaux résultats : Pour aborder ces questions, nous considérons des mécanismes de "diffusion épidémique": chaque usager choisit seul vers quel autre usager transférer des données, et quelles données transmettre. Ainsi, chaque « paquet » de données se propage dans le système comme une épidémie. Les détails du mécanisme déterminent comment ces propagations interagissent, et donc la performance du système.

Selon la première stratégie que nous considérons, chaque usager choisit vers qui transmettre au hasard, et transmet alors le plus récent paquet de données qu'il détient. Dans ce cadre, on montre qu'un délai de transmission optimal est possible, pour un débit de transmission de $1 - \exp(-1) \approx 63\%$ de l'optimal.

Ce résultat illustre le fait que des décisions prises au hasard peuvent conduire à une solution désirée. C'est l'idée de base des *algorithmes probabilistes*, dont les mécanismes de diffusion épidémique fournissent une illustration très parlante.

Selon la deuxième stratégie que nous considérons, chaque usager choisit de transmettre vers l'autre usager à qui il peut apporter le plus de données utiles, après quoi il envoie un paquet de données utile choisi au hasard. Dans ce cadre, nous montrons qu'un débit de transmission optimal est possible.

Conclusions et perspectives : Ces mécanismes simples, locaux, permettent l'utilisation efficace de toutes les ressources de communication disponibles dans un grand réseau pair-à-pair. Il semble donc ne pas y avoir d'inefficacité intrinsèque à utiliser des mécanismes complètement décentralisés. De nombreuses questions demeurent, notamment sur la relation entre débit de diffusion et délai de propagation, et sur la gestion de diffusions concurrentes.

Références :

S. Sanghavi, B. Hajek et L. Massoulié, "Gossiping with multiple messages", in IEEE INFOCOM 2007.

L. Massoulié, A. Twigg, C. Gkantsidis et P. Rodriguez, "Randomized decentralized broadcasting algorithms", in IEEE INFOCOM 2007.

Modèles et algorithmes pour la description des interactions macro-moléculaires : le triptyque biophysique - géométrie - statistiques

Frédéric CAZALS
INRIA Sophia, projet Geometrica

Cet exposé traitera des questions de modélisation et d'algorithmique en biologie structurale.

Dans une première partie, nous présenterons deux apports de la modélisation géométrique aux modèles biophysiques rendant compte de la formation et de la stabilité des complexes macro-moléculaires. Nous montrerons d'abord comment certains complexes simpliciaux associés aux diagrammes de Voronoï affines permettent de décrire précisément la structure des interfaces macro-moléculaires. Nous montrerons ensuite comment des arrangements de sphères permettent de modéliser les environnements atomiques observés dans les protéines. Dans une seconde partie et de façon plus prospective, nous expliquerons comment la biophysique, la géométrie, la topologie et les statistiques devraient permettre de mieux appréhender ces questions complexes de modélisation.

De façon transverse à l'application en biologie structurale, i.e. dans le registre algorithmique, cet exposé illustrera les difficultés posées par le développement d'algorithmes géométriques: calcul robuste - problèmes numériques, calcul efficace - problèmes combinatoires, calcul stable - quantification de l'information géométrique et topologique.

Langages formels pour la biologie systémique dans la machine abstraite biochimique BIOCHAM

Sylvain SOLIMAN
INRIA Rocquencourt, projet Contraintes

Après une brève introduction sur les nouveaux défis de la Biologie Systémique, qui doit traiter de grandes quantités de données hétérogènes sur des systèmes de complexité croissante, nous verrons comment les concepts issus de l'informatique permettent de formuler un début de réponse. En effet, des langages formels et des outils qui ont déjà fait leurs preuves pour traiter de manière modulaire des informations qualitatives et quantitatives, topologiques et multi-échelles pour les sciences de l'information peuvent tout à fait être adaptés à la modélisation du vivant, en particulier au niveau cellulaire. Nous illustrerons notre propos à l'aide du logiciel BIOCHAM, un environnement d'aide à la modélisation et à la correction de modèle qui tire parti des concepts d'abstraction et de spécification utilisés depuis des années dans l'analyse de programmes ou de circuits.