PhD Thesis: Characterizing the spatiotemporal structure and dynamics of e-science social networks

Résumé en Français

Les grilles de données et de calcul fournissent de nouveaux exemples des réseaux complexes de grande envergure qui émergent à partir d'un comportement collectif. Ces comportements sont fondés sur de multiples niveaux d'interactions. Une question intéressante est donc de savoir si ces réseaux présentent des propriétés semblables à celles des réseaux sociaux, ou bien s'en distinguent, ce qui serait la signature spécifique de l'e-science. Une conséquence opérationnelle de cette problématique est la création de modèles génératifs, associés à l'anticipation des structures des comportements. L'objectif de la thèse est la caractérisation de la structure spatio-temporelle des relations créées 1) par les co-accès aux données 2) par le trafic des tâches de calcul.

Context

Complex networks are expected to capture important characteristics of biological, social, information, and technological systems. A *complex system* consists of many interacting units, whose collective behavior cannot be explained from the behavior of the individual units alone. *Complex networks* are a special type of complex systems that can be represented with graphs whose structure is irregular, complex and dynamically evolving in time. The main focus of present research is moving from the analysis of small networks to that of systems with thousands or millions of nodes, and with a renewed attention to the properties of networks of dynamical units.

Computational grids provide new natural examples of large-scale complex networks emerging from collective behavior. Moreover, computational grids feature multiple levels of interactions. An interesting question is thus whether these networks will exhibit properties similar to those of social networks, or original ones, which would be the specific signature of e-science. An operational question is the creation of generative models appropriate for forecasting future graph structure.

Workplan

As a first step towards answering these questions, the PhD will characterize the spatiotemporal structure of the graphs created 1) by co-access to files, and 2) by the job traffic. Historical data (2008-2010) are available, and data are continuously recorded.

The requested work will cover the following three areas.

- Descriptive static analysis: the goal is to characterize the fixed-time snapshots of the trace.
- Dynamic analysis: characterize the temporal evolution of the snapshots.

For the first part, the methodology described in [1] and [2] will be followed. The characterization of self-similarity can exploit the strategy described in [3]. It can be expected that serious scalability issues might appear, offering the opportunity to an algorithmically-oriented part of the PhD: the characterization must follow the rate of graph re-shaping, thus falls in the area of streaming in the sense of real-time data mining of massive datasets.

Considering dynamic analysis, which is the core of the PhD, the context is that a pervasive characteristic of the grid is that it goes ”beyond powers laws”. To this respect, the grid behavior might be similar to many emergent natural and social systems (including the financial market): their evolution is punctuated by rare large events, which often dominate their organization and lead to huge losses (e.g in the grid software stack, catastrophic workload or request rates). Sornette proposed the concept of *dragon-kings* [4] to emphasize the importance of those events, which are beyond the extrapolation of the fat tail distribution of the rest of the population. We have the intuition that the underlying mechanism (amplification of critical cascades) has empirical reality in the behavior of the grid middleware and actors (building of collective trends), but this of course remains to explore. In any case, the important point is that the transient organization into extreme events that are statistically and mechanistically different from the rest of their smaller siblings (bifurcation) releases precursors modeled by log-periodic power laws, and has been shown to be amenable to forecasting. In a bottom-up approach, we propose to explore the applicability of these concepts to the empirical time series. In a top-down approach, we propose to do the same for a priori models of either the distributed grid actors (e.g. modeled as the participants in minority games), and to candidate middleware components (e.g. Reinforcement Learning based supervisors).

[1] Jurij Leskovec. Dynamics of Large Networks. PhD Thesis.
http://www.cs.cmu.edu/~jure/pubs/thesis/jure-thesis.pdf

[2] Adriana Iamnitchi, Matei Ripeanu, Ian Foster. Small-World File-Sharing Communities. Infocom 2004, Hong Kong, March 2004.
http://www.csee.usf.edu/~anda/papers/iamnitchi_infocom04.ps

[3] Zhou, Jiang, Sornette. Exploring self-similarity of complex cellular networks: the edge-covering method with simulated annealing and log-periodic sampling. *Physica A* 375, 741-752 (2007). http://arxiv.org/abs/cond-mat/0605676

[4] D. Sornette, Dragon-Kings, Black Swans and the Prediction of Crises, International Journal of Terraspace Science and Engineering 2:1 (2009)

Related projects

The PhD will be a member of the INRIA TAO project-team http://tao.lri.fr. The thesis is part of the Grid Observatory project http://grid-observatory.org, related to EGEE and EGI (European Grid Initiative) http://www.eu-egee.org.

Supervisor

Cecile Germain-Renaud, Professor, Laboratoire de Recherche en Informatique, Université Paris-Sud cecile.germain@lri.fr

Location

Laboratoire de Recherche en Informatique, Université Paris Sud, centre d'Orsay http://www.lri.fr