Alarm prediction via space-time pattern matching

October 12, 2015

1 Contact

Please contact Anne Bouillard, Philippe Jacquet or Marc-Olivier Buob.
`anne.bouillard@ens.fr`
`marc-olivier.buob@alcatel-lucent.com`
`philippe.jacquet@alcatel-lucent.com`

The student will be hosted by Bell Labs in Nozay (paris area) or in the joint center of Bell Labs and Inria Lincs lab in Paris.

2 Introduction

The existing networks are going to become more complex and more frequently updated. This will give the rise for more frequent alarms, some of them may lead to network disruption. Since the processes involved in the network functioning are heterogeneous, a faulty behavior of a given process may come from another process functioning on a neighbor node in the network. The real difficulty is to catch the causality when we jump over space and processes. Since frequently the processes are at different layers this is difficult for a single expert to retro-engineer the faulty behavior. The challenge here is to find the alarm correlations without *apriori* knowledge of the functioning of the processes seen as black boxes. For this the plan is to use the statistic of the alarm generation in the network in order to infer the correlations rules and finally to evaluate the probability of the imminence of a major disruption. Furthermore the retro-causal analysis of the correlations can also help of the identification of the root cause of the major dysfunction and cure them before they occur.

3 The algorithmic hard locks

If we omit the absolute and exact timing of events, the stream of alarms produced by a node can be considered as a discrete sequence of characters drawn from a finite alphabet (*e.g.* one symbol per alarm type). The problem of infering the correlation rules in a linear stream of event is an old problem, this is equivalent to find the minimal set of rules which fit with a set words of an unknown computer language. Nevertheless the exercise is heavy, *e.g.* the Coke-Younger-Kasami algorithm [1]as it takes via dynamic programing more than n^3R steps where n is the length of the stream and R is the number of rules obtained at the end. If $n = 10^4$ and $R = 100$ we get a far too expensive algorithm (need a pair of weeks on a cluster). On the other hand, very simple heuristics have been developed in [2], consisting in applying elementary transformation rules on the stream of alarms. This results in a small graph describing sharply the main alarms correlations leading to the failure of a node. Although this method have been successfully applied to logs of alarms, no formal guarantee was found ensuring the correct behavior of the algorithm. Furthermore these algorithms do not answer the question because they do not address the spatial possibility of the correlations. There is currently no language theoretic aspect which solve the space aspects of the problem since the theory is mainly focused on linear sequences. Each branching on the network is a potential choice in the causality analysis and the accumulation of branching choices may lead to a strongly exponential algorithm. The idea which consists into merging all the streams in a single one and then use the classic inference will not work well because it kills the spatial correlations as well by erasing the graph structure of the system (regardless that it dramatically increases the complexity). Basic machine learning is also difficult to implement due the exponential number of possible states coming from the same reasons.

4 Space-time pattern matching

An efficient way of depicted correlations in a linear sequence via stream statistic is the pattern matching via suffix trees. Suffix trees are data structures which are intensively used in DNA sequence analysis [3], [7]. The construction of the suffix tree is linear in n, or in $n \log n$ in incremental mode. Finding the most likely correlations is equivalent the longest matching branch in the suffix tree, which takes $\log n$. The suffix tree is also the foundation of an universal predictor algorithms over sequence [4].

The idea is to invent a new data structure which extends the suffix tree to sequences which occur in a graph. A special case is when the graph is a linear chain and in this case the data structure coincides with the suffix tree. The structure could be called the suffix GPL tree, since it could be inspired from the new structure called Graph Path Label (GPL) tree structure which has been invented in 2014 [5], [6] and is used for fast retrieval of distinct sequences which are drawn a fixed common graph. The aim is to extend the suffix tree on this structure by conveniently redefining the concept of suffix tree in a graph sequence.

5 Work program and skills

The intership should be articulated on three main direction:

1. the analysis of the complexity of GPL trie and suffix GPL tree under stochastic sequence generation models, *e.g.* Markovian models. The analysis can be held in distribution or in average (and further moments);
2. the optimisation of the GPL structure in order to prevent long matching sequence to lead to exponential explosion;
3. the design of universal predictors based on space-time pattern matching and experimentation on stochastically generated sequence and real alarm logs.

The candidate should show skills in mathematics, in particular in analytic combinatorics, information theory, stochastic processes. (S)he should be able to confront theory with full scale experiments via simulation and implementations.

References

[1] Gonzalez, R. C., Thomason, M. G. (1978). Syntactic pattern recognition: an introduction.

[2] Anne Bouillard, Aurore Junier and Benoît Ronot. Impact of Rare Alarms on Event Correlation. 9th international conference on Network and Service Management (CNSM13).

[3] Jacquet, Philippe and Szpankowski, Wojciech. Autocorrelation on words and its applications: analysis of suffix trees by string-ruler approach. Journal of Combinatorial Theory, Series A, 1994, vol. 66, no 2, p. 237-269.

[4] Jacquet, Philippe and Szpankowski, Wojciech, and Apostol, Izydor. A universal predictor based on pattern matching. Information Theory, IEEE Transactions on, 2002, vol. 48, no 6, p. 1462-1472.

[5] Jacquet, Philippe. Trie structure for Graph Sequences. Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, 2014, p. 181.

[6] Jacquet, Philippe and Magner, Abram. Variance of Size in Regular Graph Tries.

[7] Jacquet, Philippe and Szpankowski, Wojciech. Analytic Pattern Matching: From DNA to Twitter. Cambridge University Press, 2015.