Challenge 2 Results

AlgoDEEP Meeting
Rome, July 14-15, 2011
C2: Massive Data Analysis

- **Large datasets**
  - are found in biology, marketing, network analysis, program analysis, etc
  - may be *static* or *dynamic*

- **Challenging tasks**
  - extracting significant information (what is significance?)
  - identify clusters
  - understand dynamics
C2.1: Extraction of Significant Properties and Patterns

- Dynamic program analysis: efficient techniques for runtime invariant checking and error localization
- Pattern extraction from static data: more accurate characterization of significance; statistical significance of frequent itemsets; discovery of motifs that occur repeatedly

Expected results

- C2.1, M12: constraint-based framework for dynamic invariant checking based on efficient incremental change propagation algorithms. Statistical techniques to guide the mining of frequent itemsets. Extraction of long motifs from genomic sequences.

July 14-15, 2011
C2.1: Results

- **C2.1, M12**: constraint-based framework for dynamic invariant checking based on efficient incremental change propagation algorithms.

- We have designed and implemented DC, an extension of the C and C++ languages that supports one-way dataflow constraints embedded in object-oriented programs. Using this approach, we have explored program analysis applications including data structure checking and repair.

C2.1: Results

- **C2.1, M12**: statistical techniques to guide the mining of frequent itemsets.
  - We developed a novel methodology to identify a meaningful support threshold for extracting frequent itemsets which can flagged as statistically significant with a small false discovery rate.
  - We developed and analyzed new tool, called MADMX (for MAximal Dense Motif eXtraction) which extracts frequent motifs from biological sequences. We introduce the notion of density, a simple and flexible measure for bounding the number of don’t cares in a motif, to single out the “significant” motifs.
  - We introduced a computational framework for de novo identification of subnetworks in a large gene interaction network that are mutated in a significant number of patients.
C2.1: Results

- **C2.1, M12**: extraction of long motifs from genomic sequences.
  - PI proposed efficient algorithms for filtering, masking, extracting, and listing patterns from sequences, trees, and graphs.
C2.1: Results

- **C2.1, M12**: extraction of long motifs from genomic sequences.
  - PI proposed efficient algorithms for filtering, masking, extracting, and listing patterns from sequences, trees, and graphs.
C2.1: Expected Results

- **C2.1, M24**: efficient external-memory data structures for recording trails of operations. discovery of motifs with wildcards. discovery of permutation patterns.
  - pending?
C2.1: Un-Expected Results (1)

- **C2.1, extras:** further contributions not originally envisioned in the research proposal

  - Related to the new suffix-tree-like data structures:

  - Development of a progressive sampling approach for efficiently mining approximate top-K frequent itemsets, with bounds on the sample size as function of the desired approximation:

  - Development of an optimized data structure for high-throughput 3D proteomics data:
C2.1, extras: further contributions not originally envisioned in the research proposal

- We highlighted a fundamental problem in using the (N)PRI metric to assess the quality of segmentation of melanocytic lesions:

- Related to the extraction of significant properties and patterns:
  - D. C. D'Elia, C. Demetrescu, I. Finocchi, Mining Hot Calling Contexts in Small Space, 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'11), ACM Press, 516-527,
C2.2: Analysis of Dynamic Networks

- Analysis of routing and traffic information
- Visualization and exploration tools

Expected results

- **C2.2, M12**: large repository of routing updates and RIB dumps. algorithms to identify hot-spots, understand routing dynamics, extract structural properties and clusters. visualization of stream of edges. visualization of non-planar dense subgraphs and clusters.

- **C2.2, M24**: usage of information gathered to optimize and reduce instability of BGP and detecting causes of instability in real-time. implementation and experimental validation of clustering algorithms. tools to visualize clustered graphs and stream of edges. algorithms for the visualization of non-planar networks.
C2.2: Results

- **C2.2, M12**: large repository of routing updates and RIB dumps.
  - A large repository of routing updates now exists at RM3, collecting the updates of several months of year 2010 and of year 2011.
  - Datasets (topology and traffic information) to be employed in the experimental validation of clustering algorithms
    - Large datasets are collected by RM2
C2.2: Results

- **C2.2, M12**: algorithms to indentify hot-spots, understand routing dynamics, extract structural properties and clusters.

- **Tools and algorithms for mining structural properties of the AS network:**

- **Monitoring the evolution of the number of prefixes traversing each edge:**

- **Clustering algorithms capable of extracting clusters from the information contained in the traffic matrix.**
C2.2: Results

- **C2.2, M12**: algorithms to identify hot-spots, understand routing dynamics, extract structural properties and clusters.

- In this task we studied several algorithmic techniques for analyzing structural properties of networks. Addressed problems include finding articulation points and bridges in directed networks, as well as detection of real-time anomalies. As a relevant application scenario, our techniques have been applied to the network of Autonomous Systems of the Internet.


C2.2: Results

- **C2.2, M12**: visualization of streams of edges. visualization of non-planar dense subgraphs and clusters.
  - theoretical area bounds for the visualization of streams of edges and algorithms to handle the cases of paths and trees.
  - bounds for the identification of dense clusters (cliques, k-core-graphs, k-connected, etc) producing highly readable graphs such as trees, planar graphs, etc.
C2.2: Results

- **C2.2, M12**: visualization of non-planar dense subgraphs and clusters (1/3).

  - new visualization paradigms for non-planar graphs, with particular attention to the area requirement and to the angles at which edges cross.
C2.2: Results

C2.2, M12: visualization of non-planar dense subgraphs and clusters (2/3).

• Results on simultaneous embeddings:
  - P. Angelini, G. Di Battista, F. Frati, M. Patrignani, I. Rutter, Testing the Simultaneous Embeddability of Two Graphs whose Intersection is a Biconnected Graph or a Tree, Workshop on Combinatorial Algorithms (IWOCA '10), Lecture Notes in Computer Science, vol. 6460, 212-225, 2011.
C2.2: Results

- **C2.2, M12**: visualization of non-planar dense subgraphs and clusters (3/3).

  - Results on monotone and greedy drawings:

  - Algorithms for the representation of a clustered network where edges are straight-lines and each cluster is represented by a box.
    - P. Angelini, F. Frati, M. Patrignani, Splitting Clusters To Get C-Planarity, 17th International Symposium on Graph Drawing (GD '09), 2010.
C2.2: Expected Results

- **C2.2, M24**: usage of information gathered to optimize and reduce instability of BGP and detecting causes of instability in real-time. Implementation and experimental validation of clustering algorithms.
  - Implementation of the cluster visualization algorithms developed during the first year.
    - VHYXY: Visual Hybrid (X,Y)-clustering (VHYXY). System for the visual analysis of social networks. It has been mainly designed to visualize co-authorship networks, and it has been experimented on data from DBLP.
    - The system implements new clustering algorithms based on k-cores and hybrid visualization techniques that combine node-link and matrix-based representations.
C2.2: Expected Results

- **C2.2, M24**: tools to visualize clustered graphs and streams of edges. Algorithms for the visualization of non-planar networks.
  - tools for the visualization of streams of edges
    - two independent implementations (RM1 and RM3 theses)
  - tools for the exploration of large networks that use hybrid visualization techniques featuring an automatic identification of the portions to be succinctly represented.
    - VisFan: Visual Analysis of Financial Networks. System for the visual analysis of social networks coming from financial data. It provides a support to those analysts that have to discover financial crimes, like frauds and money laundering. The system implements new interactive techniques for handling clustered graphs.
C2.2: Un-Expected Results

- **C2.2 extras**: further contributions not originally envisioned in the research proposal

  - RM2 jointly with RM1 contributed to algorithmic techniques for studying connectivity properties such as articulation points and bridges in directed graphs:
  - We showed that the well-known PageRank algorithm is extremely sensitive to the tiniest variations in the damping factor, at least on some pathologic graphs:
    - M. Bressan, E. Peserico. “Choose the damping, choose the ranking?” J. Disc. Algs. 8(2) 2010
  - Analysis of dynamic networks
## Units Involved

<table>
<thead>
<tr>
<th></th>
<th>C1.1</th>
<th>C1.2</th>
<th>C2.1</th>
<th>C2.2</th>
<th>C3.1</th>
<th>C3.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Padova</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Pisa</td>
<td>+</td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Roma 1</td>
<td></td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Roma 2</td>
<td></td>
<td>+</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roma 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The symbol * denotes direct involvement
- The symbol + denotes a mere interest