Interactive Visualization of Provenance Graphs for Reproducible Biomedical Research

Stefan Luger, Holger Stitz, Samuel Gratzl, Nils Gehlenborg and Marc Streit

JOHANNES KEPLER UNIVERSITY LINZ

HARVARD MEDICAL SCHOOL

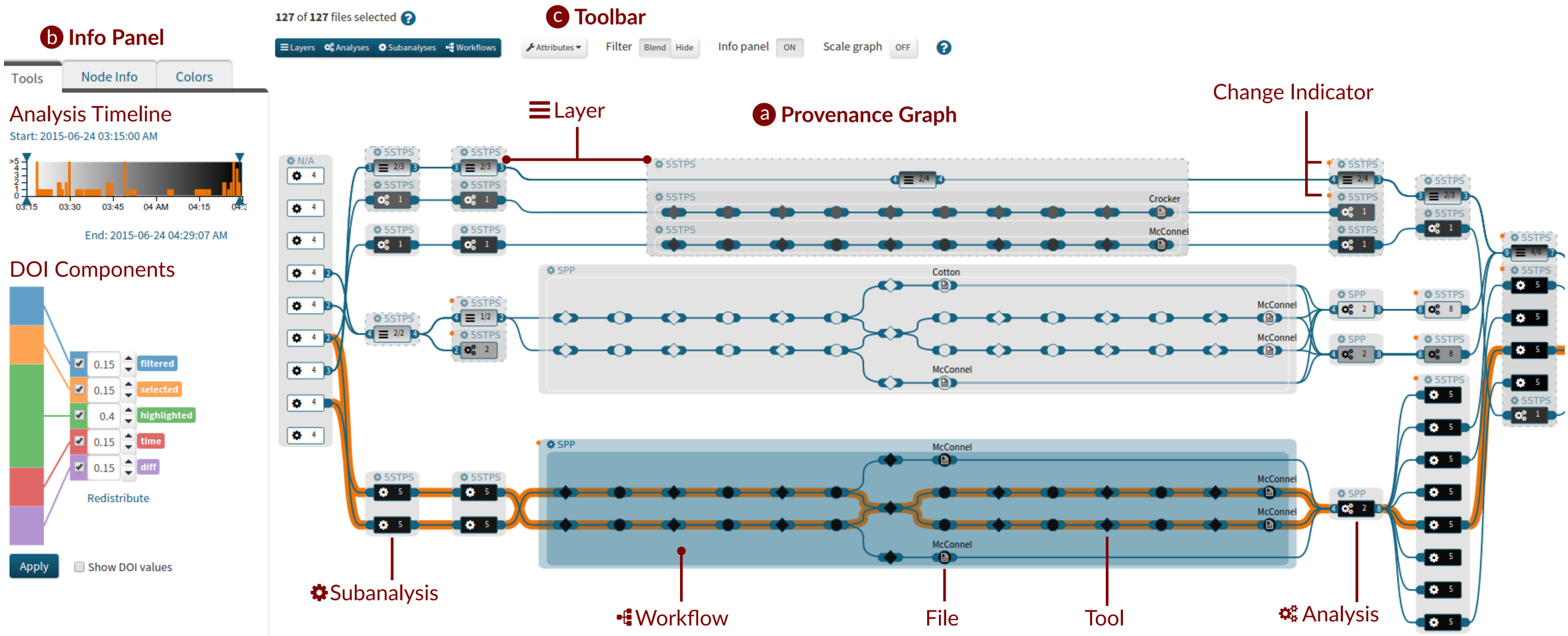

Figure 1: The provenance graph (a) is aggregated and filtered based on the selected workflow execution time and the weighted degree-of-interest (DOI) components (b). The ***Node Info*** tab provides details-on-demand while the ***Colors*** tab let users define a custom color scheme. In the top center of the graph (a), two horizontally aligned workflows show a compound layer node, where the top node represents the layer itself while two workflows are extracted based on their specific DOI exceeding a predefined threshold. The toolbar (c) provides node type specific views (Fig. 3) and attribute mapping to nodes.

The *Refinery Platform* (Fig. 2) is an integrated data management, analysis, and visualization system designed to support reproducible biomedical research. Refinery collects provenance information about biomedical workflows applied to heterogeneous datasets in large genomic studies. We present a visualization technique that dynamically reduces the complexity of subgraphs through hierarchical aggregation and application of a degree-of-interest (DOI) function to each node. We further reduce complexity of the provenance graph visualization by layering identical or similar sequences of parallel analysis steps into an aggregated sequence.

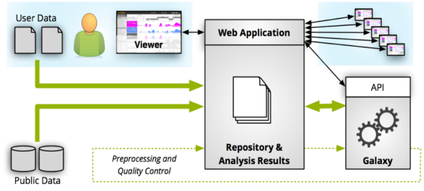

Figure 2: Refinery Architecture

Modular Degree-Of-Interest Function To determine the current user interest on any node in all hierarchy levels, we use a modular degree-of-interest (DOI) function [1,4]. The weighted DOI function incorporates multiple components (Fig. 1b):

- properties of the graph (e.g., date and time, changes over time)
- interest derived from user actions such as filtering, node selection, and highlighting

The DOI computed controls the degree of hierarchical aggregation to the nodes.

Most existing approaches (e.g, [2]) are based on node-link diagrams that do not scale well to large graphs. From a visualization research point of view, provenance graphs comprise two major **challenges:**

- they become large very quickly and
- they contain time-dependent information.

With domain experts **5 tasks** were elicited:

- **High-Level Overview:** Present general information about workflow runs, type and time
- **Attribute Encoding:** Node glyphs encode attributes such as type or creation date with visual channels
- **Drill-Down on Demand:** Manual and automatic control of hierarchy levels to show details on demand
- **Investigate Changes:** Communicate changes over time in every hierarchy level
- **Investigate Causality:** The chain of files and transformations that led to a particular result

[1] J. Abello, S. Hadlak, H. Schumann, and H.-J. Schulz. A Modular Degree-of-Interest Specification for the Visual Analysis of Large Dynamic Networks. IEEE Trans. on Visualization and Computer Graphics, 2013.
[2] J. Freire and C. T. Silva. Making Computations and Publications Reproducible with VisTrails. Computing in Science & Engineering, 2012.
[3] E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen. Visual Compression of Workflow Visualizations with Automated Detection of Macro Motifs. IEEE Trans. on Visualization and Computer Graphics, 2013.
[4] F. van Ham and A. Perer. "Search, Show Context, Expand on Demand": Supporting Large Graph Exploration with Degree-of-Interest. IEEE Trans. on Visualization and Computer Graphics (InfoVis '09), 2009.
[5] H. Stitz, S. Gratzl, S. Luger, N. Gehlenborg, and M. Streit. Transparent layering for visualizing dynamic graphs using the flip book metaphor. In Poster Compendium of the IEEE VIS Conference. IEEE.

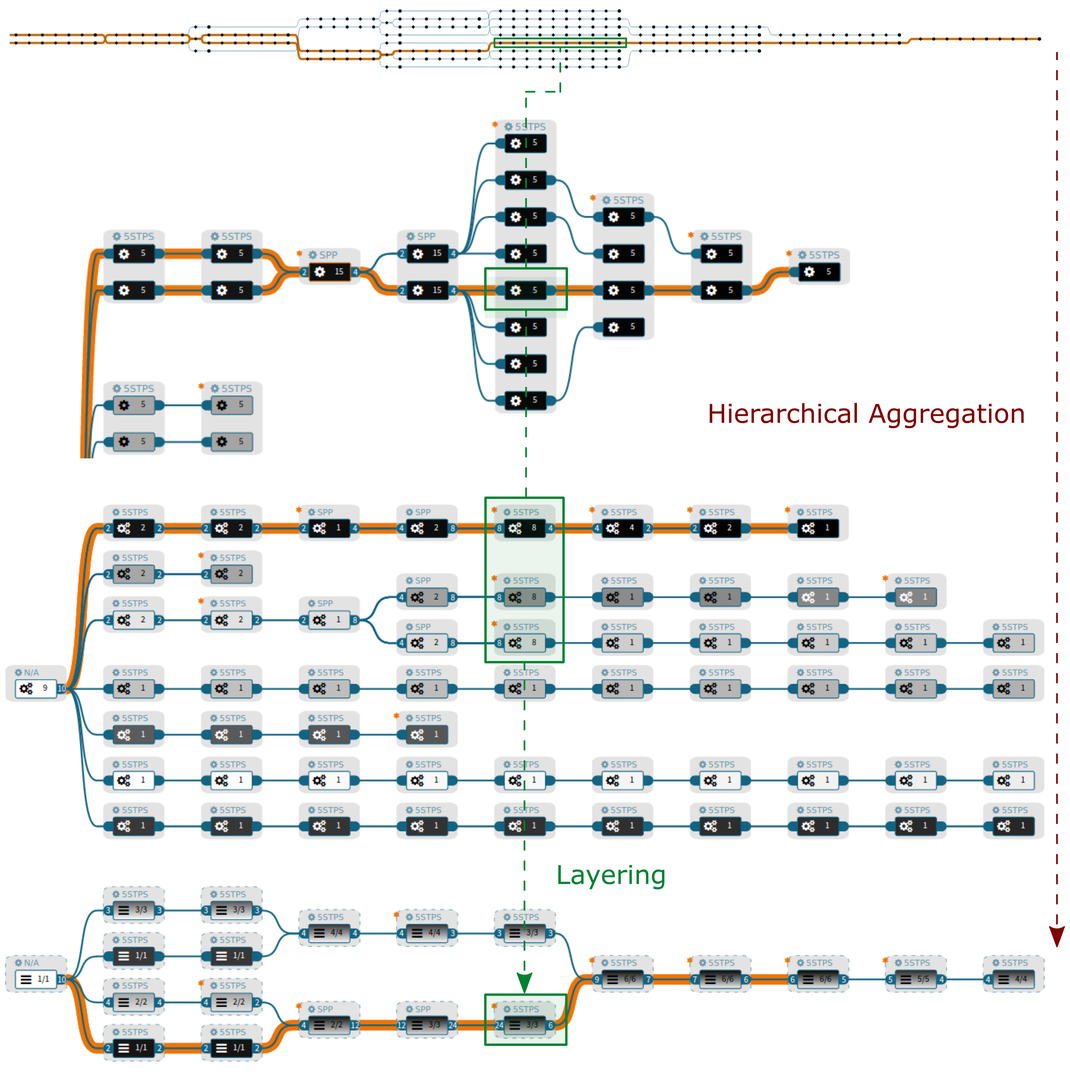

Figure 3: The graph in Fig 1. is shown in each of the **4 hierarchy levels**. In Refinery the provenance graph consists of analyses (c), which in turn consist of subanalyses (b) that represent a workflow execution on a set of input files. The example graph contains 1100 files/tools, 100 subanalyses, and 60 analyses.

(a) Workflow (Level 0)
Files and tools represent the atomic building blocks of a workflow.

(b) Subanalysis (Level 1)
A subanalysis is restricted to exactly one workflow template.

(c) Analysis (Level 2)
An analysis usually contains multiple subanalyses where the same workflow is executed on a combined set of input files.

(d) Layer (Level 3)
We use network motif discovery [3] to detect and aggregate similar analysis paths into a compound layer node.

A motif is constrained to:

- workflow type
- parameters
- subanalysis count
- in- and outgoing edges

Contact
lug.stefan@gmail.com
https://github.com/sluger

REFINERY PLATFORM
http://www.refinery-platform.org/

Acknowledgements
Austrian Research Promotion Agency (FFG) (840232),
Austrian Science Fund (FWF) (P27975-NBL), and
US National Institutes of Health (K99 HG007583)

FFG

FWF Der Wissenschaftsfonds.