CroP

Coordinated Panel Visualization

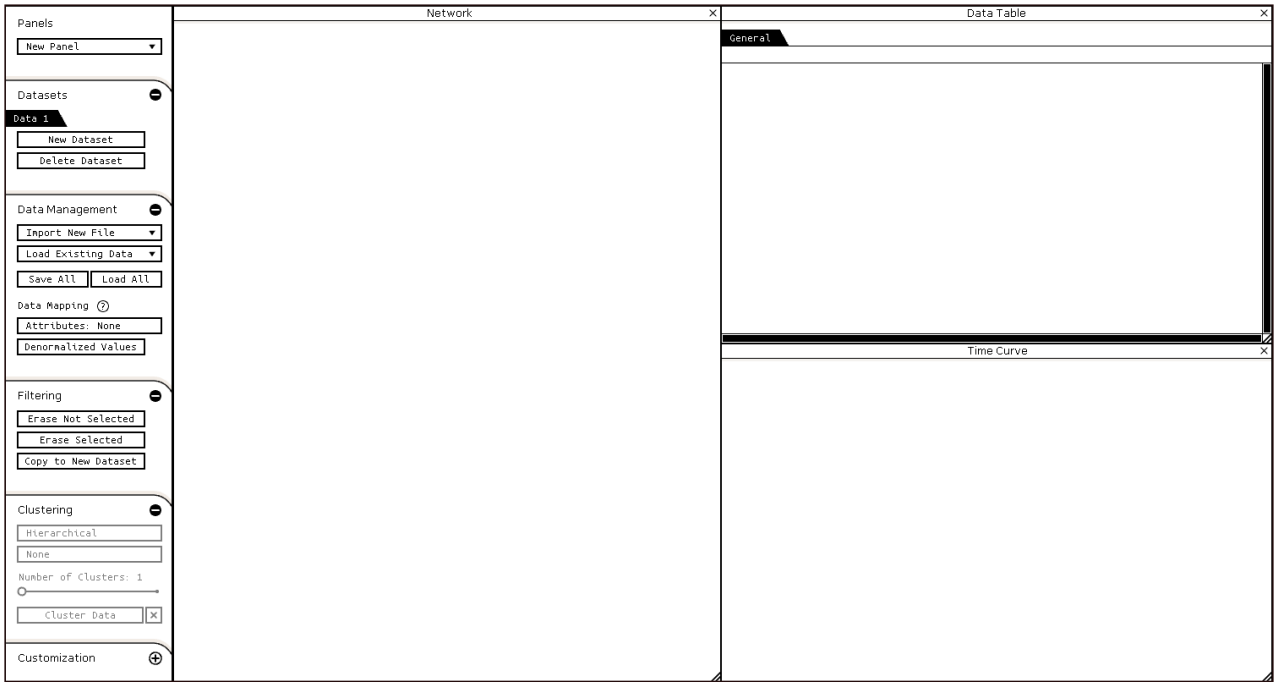
User Manual

CroP is a visualization tool for visualizing and analyzing networks integrated with temporal or multivariate data in order to identify patterns and behaviors that happen over time. Loaded data can be visualized through dynamic models within flexible panels, where interactions are coordinated between them for the same datasets.

1. Initialization

CroP opens with a sidebar of options on the left and several default visualization panels occupying the remainder of the screen. The size of CroP's window can be adjusted be dragging the edges or maximizing it, which will adjust the size of the panels accordingly. Some options in the sidebar are hidden by default but can be toggled by pressing the button on the top right of each one.

To begin visualizing data, datasets files can be uploaded into the program by selecting the "Import New File" dropdown within the "Data Management" section in the options sidebar, selecting the type of file you wish to load, and then choosing the file from your computer.



Default panel configuration of CroP, before data is loaded.

CroP offers some support for analyzing biological datasets by cross-referencing names of loaded data points with an integrated Gene Ontology database. Additionally, existing databases can be loaded from the "Load Existing Data" dropdown, which contains a human protein-protein interaction network.

1.1. Supported Files

The application is able to receive network data, time-series data, and variables data. These data files can be formatted as either comma-separated values (CSV) or tab-separated values (TSV). The following examples use CSV:

Network data – Files should consist of a list of edges, one per lines. Each edge must contain two attributes: the name of the first node and the name of the second.

```
Node 1, Node 2
Node 1, Node 3
Node 1, Node 5
Node 2, Node 3
Node 2, Node 4
(...)
```

Time-series data – File should contain the name of each node followed by an ordered list of values corresponding to each time point. Additionally, the first line of this file may contain a list that defines the name of each time point (such as the hour or date).

```
Metric, Time 1,  Time 2,  Time 3,  Time 4,  ...
Node 1, Value 1, Value 2, Value 3, Value 4, ...
Node 2, Value 1, Value 2, Value 3, Value 4, ...
Node 3, Value 1, Value 2, Value 3, Value 4, ...
Node 4, Value 1, Value 2, Value 3, Value 4, ...
(...)
```

Variables data – File should contain the name of each node followed by an ordered list of values corresponding to each variable, which are specified in the first line. This file type should be chosen when dealing with multiple values that do not have a defined order (unlike time-series).

```
Free Space, Var 1,   Var 2,   Var 3,   Var 4,   ...
Node 1,     Value 1, Value 2, Value 3, Value 4, ...
Node 2,     Value 1, Value 2, Value 3, Value 4, ...
Node 3,     Value 1, Value 2, Value 3, Value 4, ...
Node 4,     Value 1, Value 2, Value 3, Value 4, ...
(...)
```

If any values in time-series and variables files are left blank, the application will interpret this as nodes being "inactive" at those time points or for those variables.

When multiple files are loaded, options will be presented to either merge or filter data. If a node has the same name on different types of files, this will merge all the data into a single node. The application will also ask if you would like to filter new nodes that don't correspond to existing nodes or vice-versa, if you would like to only keep nodes that match every type of data.

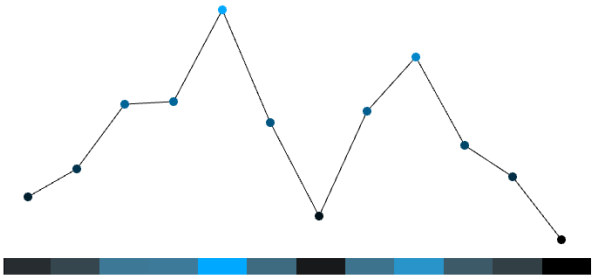
Selecting the "Save All" option will generate a file containing the current state of CroP, which includes the entire dataset as parsed by the application, clustering, panels settings and parameters. This file can be loaded at any point through the "Load All" option in order to restore your workspace and all of the data you were working on, with minimal loading times.

1.2. Basic Representation

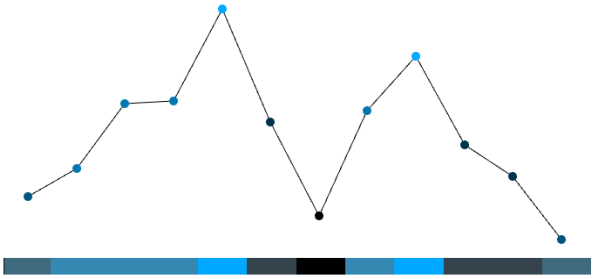
As default, time-series and variable values are mapped from black to blue, where black represents the lowest data value, and bright blue represents the highest. Time-series data can also be mapped by variation or tendency. The type of mapping can be changed in the options sidebar, within "Data Management" and under "Data Mapping". Mapping uses normalized values by default, but they can be denormalized within the previous options.

Variation mapping shows how values shift from time point to time point, representing black as the strongest decreases, bright blue as the sharpest increases. Colors present grayer tones when values didn't change significantly from the previous time point.

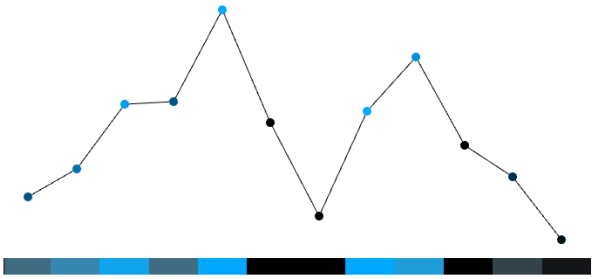
Tendency mapping shows how values shift between the previous and next time points, representing peaks of data as bright blue and valleys with black. Continuous increases are presented with brighter colors while continuous decreases are presented with dark tones. This mapping is primarily used to identify shifts in tendency in specific datasets.



Example of value color mapping.



Example of variation color mapping.



Example of tendency color mapping.

Other default colors include temporal progression being mapped from black to orange, inactive nodes being represented with red hues, and values being compared across different datasets with green hues. These colors can be changed in the Customization section of the options panel.

1.3. Filtering Data

Portions of the dataset can be removed or shifted using the options provided in the sidebar under “Filtering”. The button “Erase Not Selected” will remove every data point that is not currently selected in any of the visualization panels, while “Erase Selected” will remove those that are selected. Selecting data points in each type of panel is covered in the “Visualization Panels” section.

The button “Copy to New Dataset” will instead create a new dataset containing only the selected data points, without removing them from the current dataset. Managing multiple datasets is further explained in the “Comparing Multiple Datasets” section.

1.4. Clustering

The options located in the “Clustering” section in the options panel can cluster data into groups of nodes with similar proprieties. These options select the type of clustering, the attribute being clustered, and the merging criteria. Data elements can be clustered by their position in the network panel, by temporal attributes, or by the values of variables.

Clustering may take a while depending on the size of the data, but these loading times can be bypassed after the first time for the same dataset by using the “Save All“ and “Load All“ options. After clustering, the number of clusters can be switched by using the bottom slider without needing to cluster the data again.

2. Visualization Panels

The visualization panels are located to the right of the options panel, and they will display the loaded data from files. These panels can be moved, resized or removed, while new panels can be added by selecting them from the “New Panel” dropdown in the options sidebar, under “Panels”.

Drag the top bar of a panel to move it, close it using the button on its top-right, or resize it by dragging the panel’s bottom-right corner. If other panels are overlapped when moving or resizing, those panels will be adjusted if there is free space, or removed. There is also a dropdown that selects with dataset is being visualized, which is further explained in the “Comparing Multiple Datasets” section.

2.1. Data Table

The data table is the visualization that shows every data point at its lowest level, listing them in a sortable table where rows can be selected to access the proprieties of each individual element, or the aggregated proprieties of groups.

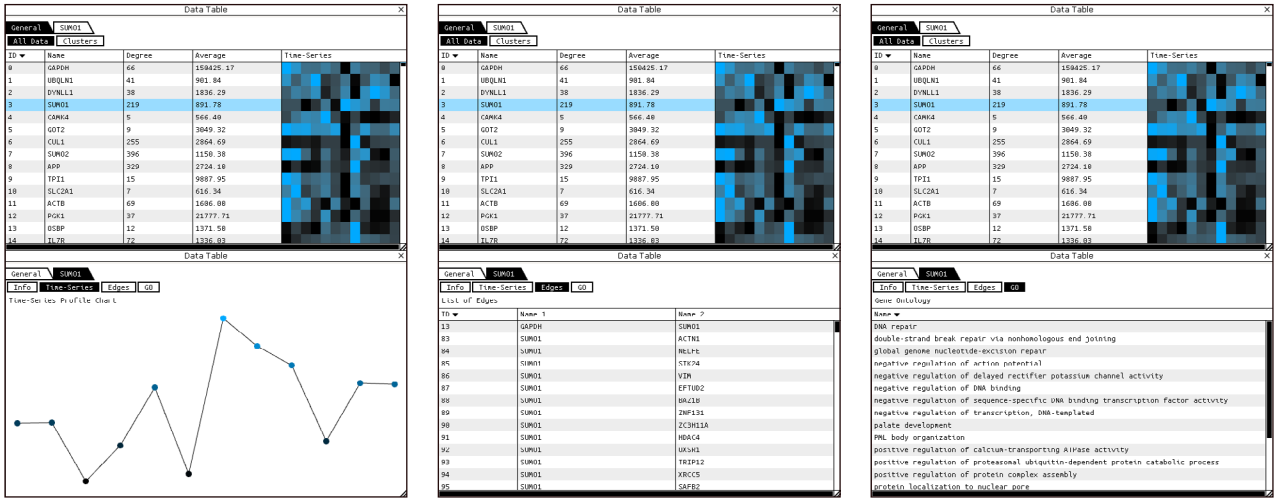
2.1.1. Selecting Rows

Tables can can be scrolled by using the mouse wheel or by clicking or dragging the sliders on its side. Elements can be selected in the data table by clicking on their respective row, and deselected by clicking them again. The positions of selected elements are highlighted in the scrollbar with colored bars. Selections are also coordinated across visualization models, highlighting the selected data elements in all network and table visualizations of the same datasets.

For each selected data point, a new tab on the top of the table is created. These contain any existing proprieties of the data point, including a line chart of its temporal profile, a list of its edges between other nodes, and a table of corresponding Gene Ontology terms. If the number of tabs exceeds the width of the panel, tabs can be dragged left and right to bring others into view.

Holding the ‘CTRL’ key will allow the selection of multiple rows simultaneously, each creating an individual tab for the selected element. Alternatively, to select

large sections of data points, holding the ‘SHIFT’ key will toggle start/stop points for row selection: selecting two separate rows in this manner will automatically select every row between them. Rows selected in this manner do not create individual tabs for every data point selected, and instead create a single tab called “Highlighted” which contains aggregated data on every data point selected.



Sets of table panels showing an element being selected (top) and various information about that element (bottom): a temporal profile (left), a list of its edges (middle), and a list of its Gene Ontology proprieties (right).

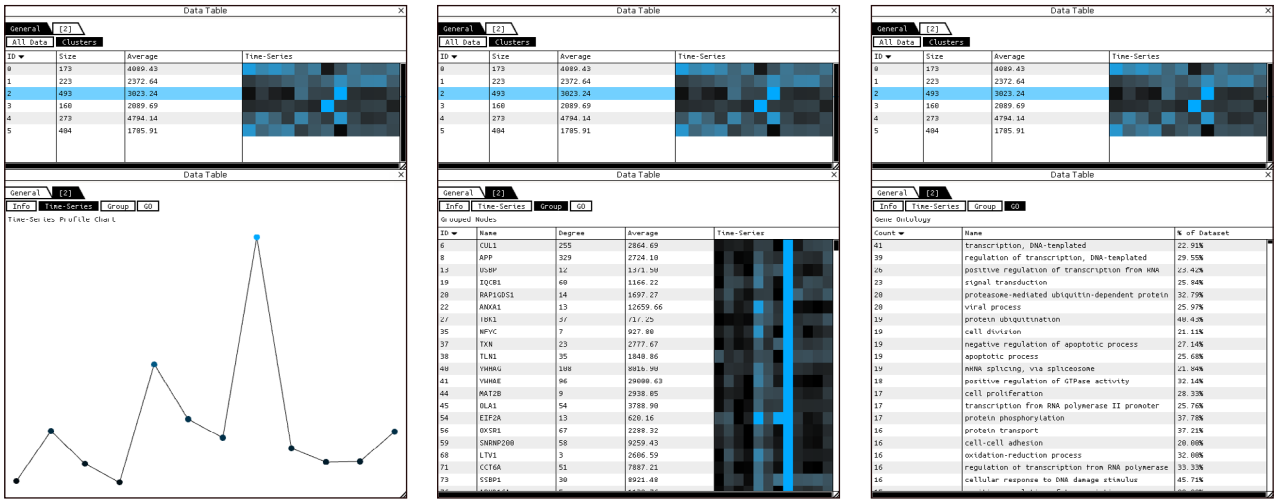
2.1.2. Sorting by Columns

The table’s columns are kept updated with the current data, meaning that loading new attributes will also add new proprieties as columns. Clicking on a column’s title sorts the table by its respective propriety and selecting the same column again will invert the current order.

If data contains temporal attributes or variables, columns will be added with a sequence of colored squares that match the respective values of each data points. These columns can only be sorted after data has been clustered as it orders nodes by similarity.

2.1.3. Other Tables

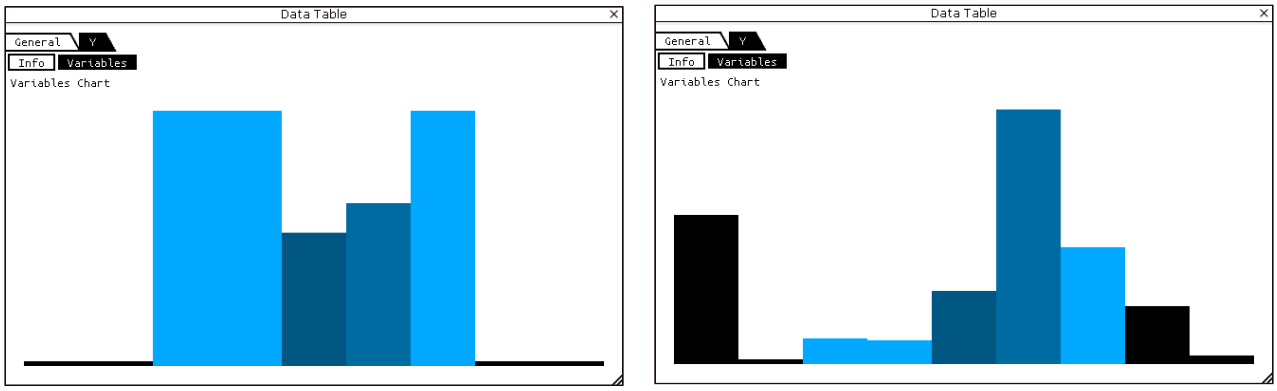
If the data is clustered, a “Clusters” button will be placed at the top of the panel, next to “All Data”. Clicking it will make the table show a list of clusters instead of all data, along with the mean values for time-series and variables of their respective nodes. Selecting a cluster will also create a tab with detailed proprieties, including a table of its nodes and a table of aggregated Gene Ontology terms.



Sets of table panels showing a list of clusters with one being selected (top) and various information about that cluster (bottom): an aggregated temporal profile (left), a list of its elements (middle), and a list of the Gene Ontology proprieties within the cluster (right).

If variables data has been loaded, a “Variables” button will be added to the top of the panel. Clicking it will create a table of all the variables and their values across the dataset.

If more than one dataset has been created, there will be a “Datasets” button at the top of the panel, which will create a table that lists all existing datasets. Using this table is further explained in the “Comparing Multiple Datasets” section.



Variable profile, normalized (left, where bar height compares between the element's various proprieties) and denormalized (right, where bar height is based on the value of each variable in relation to every other element).

2.2. Network Visualization

The network panel represents each data point as a circle in two-dimensional space. They are initially displayed in a spiral layout, in the order loaded from the file. At the top left corner of the panel, there will be options to change how the nodes are positioned based on the proprieties loaded.

The transparency of edges is mapped to their quantity on screen, where higher amounts of edges increase their transparency until they might not be drawn at all. This is done to reduce visual noise in complex networks and highlight node relationships using their position. However, selecting the "Always Show Edges" button will override this condition.

2.2.1. Layouts

The options menu provides two additional layouts for the network. The Yifan Hu layout will sort the nodes based on their edges, positioning them so that related nodes will be closer to each other. Three parameters of the layout are mapped to sliders in the options menu which can be switched to balance the calculation time of the layout against its accuracy:

Step Ratio - The ratio used to update the step size across each iteration of the calculation of the layout; higher ratios indicate a faster convergence but lower accuracy.

Quad Tree Maximum - The maximum value to be used in the quadtree representation; greater values increase accuracy at the cost of execution time.

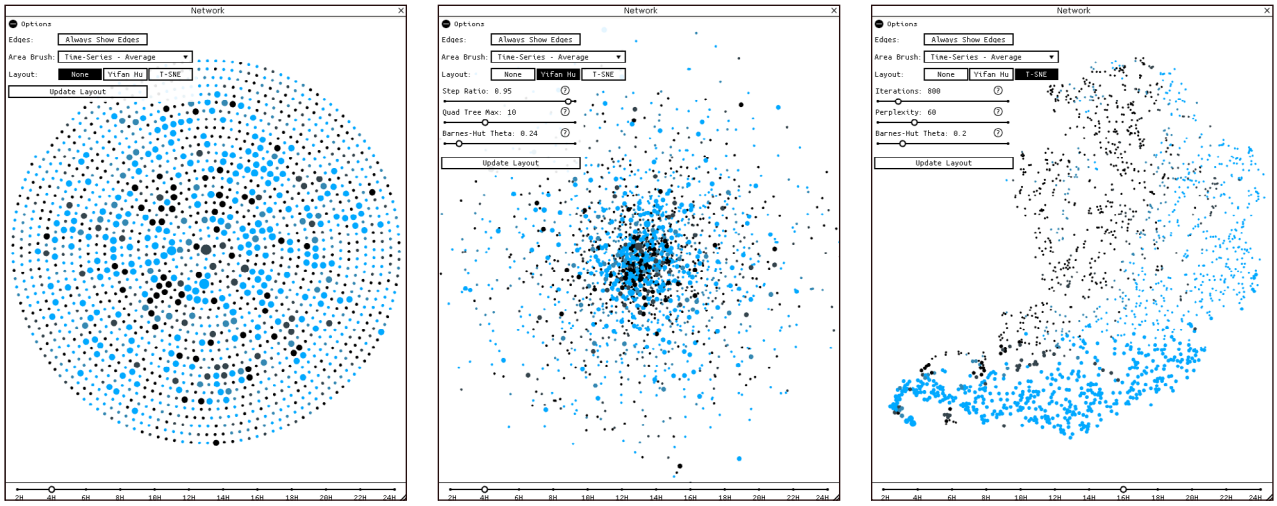
Barnes Hut Theta - Lower values increase accuracy at the cost of execution time.

The t-SNE layout will only be available if time-series data or variables data is available, as it will position nodes based on their similarity between their values, rather than edges. This is similar to clustering but without necessarily creating well-defined clusters. Three parameters of this layout can also be changed in the options menu through sliders:

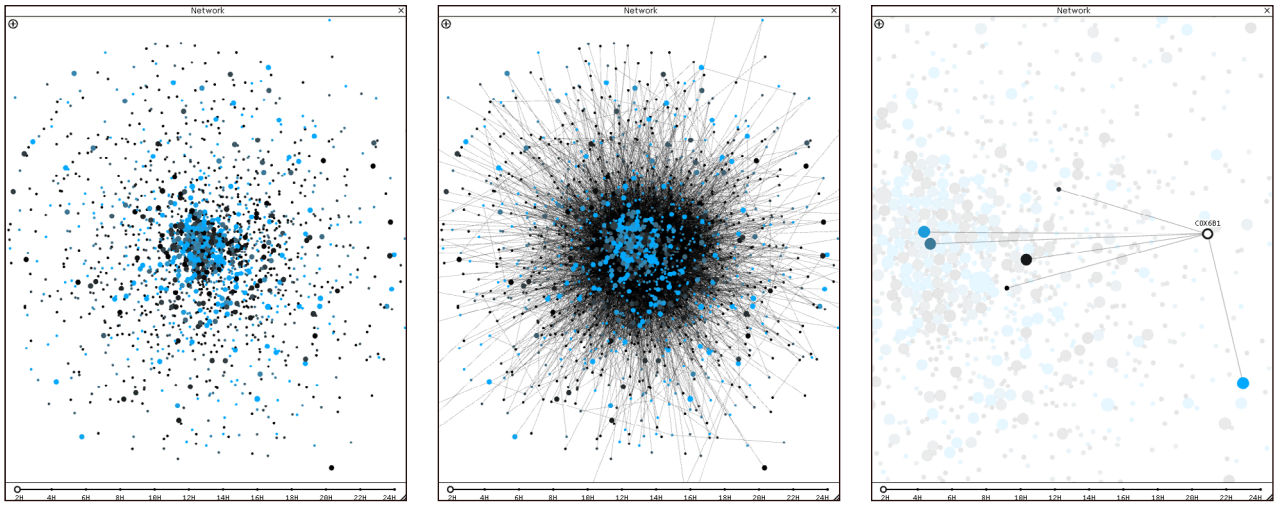
Iterations - Number of iterations performed; higher values increase accuracy at the cost of execution time.

Perplexity - Approximate guess on the number of neighbors for each point; higher values help preserve global structure, but may obscure local structure and increase execution time.

Barnes Hut Theta - Lower values increase accuracy at the cost of execution time.



Network panel sorted into three different layouts: space-filling spiral (left), Yifan Hu (middle), t-SNE (right).



In high amounts, edges are hidden by default (left), but can be shown by selecting the “Always Show Edges” button (middle). Selecting a node will only highlight that node’s edges (right).

2.2.2. Navigation

The network visualization can be panned by clicking the left mouse button and dragging the mouse, or zoomed in/out on the current mouse position using the mouse wheel.

If either time-series data or variables data have been loaded, nodes will be colored and sized according to the type of values selected (in “Data Mapping” under “Data Management” in the options sidebar). This will also create a slider at the bottom of the panel that will either display a timeline of the time-series data, or a list of all the variables. Dragging the slider will change the time point or variable that the nodes are currently mapped to, updating their size and color dynamically.

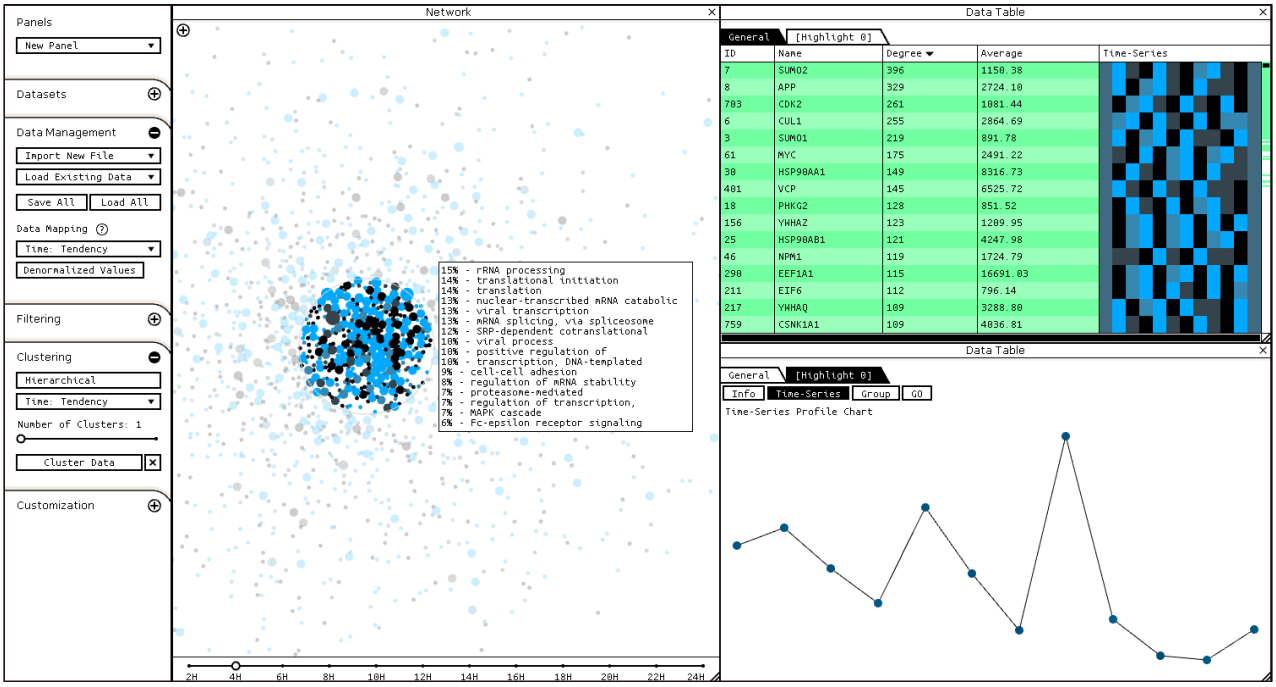
2.2.3. Selecting Nodes

Nodes can be hovered and selected with the mouse in to highlight them and to those which are directly related to them. Any selections will also be coordinated with other network and data table panels that are showing the same dataset. Clicking outside of a node will reset all selections.

If time-series or variables data have been loaded, or if names of biological nodes correspond with the Gene Ontology database, then an additional analysis tool will be available. Right-clicking anywhere on the network panel will create a circle around the mouse. This circle acts as a lens which selects every node inside of it and then displays a small visualization next to it that describes them.

For time-series and variables data, the visualization will be of a line chart that depicts an average of all the values for every node selected. For biological nodes, the lens will show the percentage of each gene ontology propriety that exists within the selected group. If there exist multiple types of data that can be represented within the lens, the preferred type of data can be chosen in the options menu. The size of the lens can be increased or decreased with the mouse wheel, and the lens can be turned off by right-clicking a second time.

Additionally, by holding the 'CTRL' key, multiple nodes can be selected, either by clicking them, or by brushing them with the lens.



Selection of a section of the network with the mouse lens, showing a count of Gene Ontology proprities. Selected nodes are also highlighted in the data table panel.

2.2.4. Data Clusters

Clustering the data will group nodes into the selected number of clusters. The position of the clusters reflects the relationships between their nodes, as related clusters will be placed closer to each other. Like nodes, clusters can also be hovered and selected by placing the mouse on their borders.

Nodes will also be sorted within each cluster in a spiral layout. The order of the nodes within each spiral is based on the overall clustering, meaning that neighboring nodes may also be more similar to each other.

2.3. Time Curve

The time curve panel provides visualizations and tools meant for the analysis of time-series datasets. A time curve consists of a timeline that is bent so that its time points are placed relatively to each other according to their similarity. As such, time points that are closer to each other represent moments at which the same data points have similar values. This may reveal patterns of how the data behaves over times, such as revealing cy cles or highlighting instances of significant changes.

2.3.1. Timeline Presentation

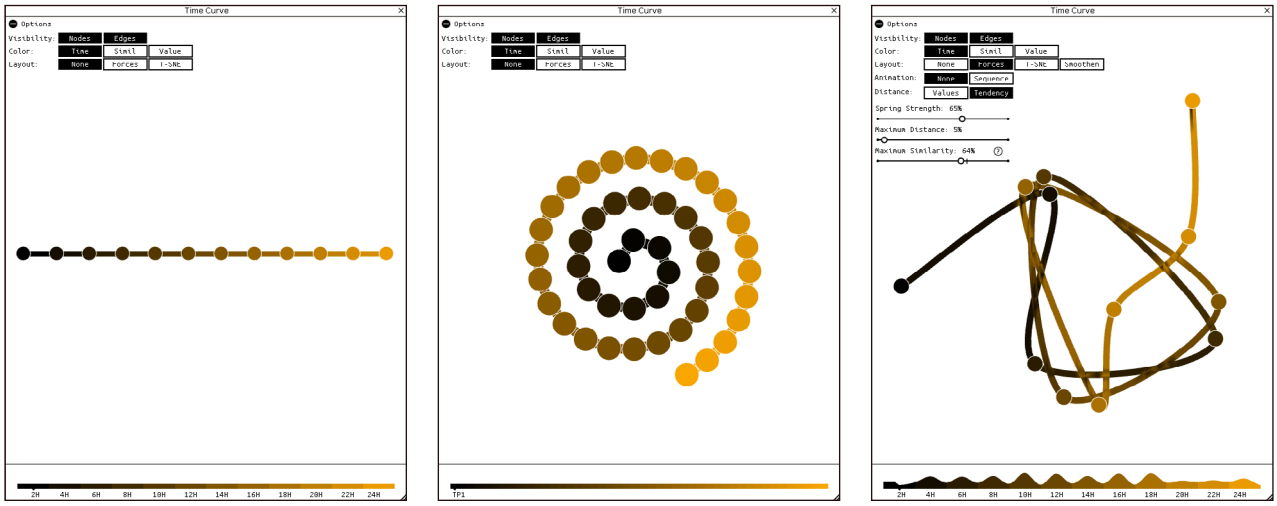
When temporal data is loaded into CroP, each time point is converted into a node and displayed sequentially as a timeline in this panel. The nodes will either be displayed as a horizontal line or as a spiral, with the latter being used when there is a large quantity of time points.

The default color of each node ranges from black to orange, indicating time progression, where black represents the initial time point and orange the final one. Each node can also be clicked it order to select it, which will also change the current time point of existing network visualizations. The mouse can also be dragged to pan the visualization and the mouse wheel will zoom in/out of the current mouse position.

The top left of the panel contains a list of options that control both visual elements and the layout of the nodes. Under “Visibility”, nodes or edges can be hidden in order to portray only the other. The “Color” options can change the default color

scheme from mapping temporal progression to mapping similarity. When this option is chosen, selecting time points will highlight all the time points that are similar to those chosen, regardless of the layout. The “Animation” options available will depend on the layout chosen but they create different animated flows between time points. The “Layouts” option will change the position of time points, and are discussed in the following two sections.

The bottom of the panel contains a slider with all the time points in sequence. When the slider is dragged, a line will be drawn across the respective time points in the above visualization. This can be used to keep track of the temporal order of the data as different layouts are applied.



The initial layout will either be linear (left) or a spiral (middle), depending on the number of time points. This layout can then be bent into a time curve using a force-directed layout (right).

2.3.2. Bending Time

Initially, the timeline visualization can be distorted using one of two layouts: Forces or T-SNE. The forces layout is dynamical, pulling similar nodes closer while repelling those that are different based on the chosen parameters.

Spring Strength - Increases the speed at which nodes move into positions that reflect their similarity, but high values may also cause instability.

Maximum Distance - Determines the maximum distance between the most dissimilar nodes; increasing this value will expand the size of the visualization.

Maximum Similarity - This slider maps the percentage of the dataset that must be similar between two time points in order for them to be at maximum proximity. This means that at 50%, time points that are very close to each other should have at least 50% of data points with similar values. If there is a mark on the slider, it indicates the highest similarity value found between the current time points.

The t-SNE layout also positions nodes based on the chosen distance metric. However, unlike the previous layout, it is not dynamical, being calculated only when the "Update" button is pressed. Additionally, it is also not deterministic, meaning that each recalculation of the layout will yield different results. The quality of these results can be influenced by the parameters, which balance accuracy and processing time. While each of these parameters is explained by hovering over the (?) icon, they are also addressed in Section 2.1.2.

Compared to t-SNE, the forces layout is capable of more easily discovering positions that best reflect the relative similarity between time points, but may present issues with sorting large quantities of nodes. As such, the T-SNE layout may yield faster and better results when dealing with complex time series.

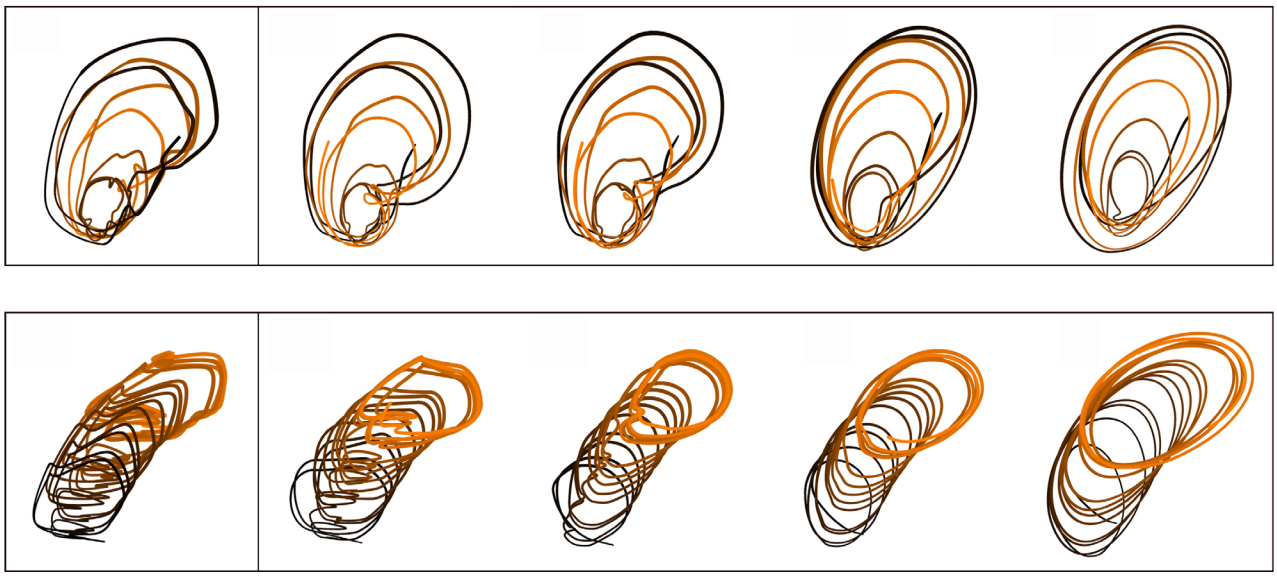
2.3.3. Smoothening

After applying either of the previous layouts, the Smoothen layout option will become available. This layout is applied on top of the previous one, and redraws the time curve visualization to create smoother edges with gradual transitions between time points. This layout adds two new animation options that represent the flow of time with either pulsating edges, or travelling arrow particles.

While the default parameters should create a smoother visualization, these can also be altered to further distort the timeline and create new types of visualization. There are two parameters that can be shifted:

Smoothness - Affects the distortion between each time point; increasing this value will create more continuous shapes and reduce perturbations caused by smaller shifts in the data.

Intermediate Points - Defines the number of points that make up the edges drawn between time points; increasing this value increases how well-defined each edge is drawn, but extreme values will cause distortions.



Various visualizations of two datasets created through the smoothening layout by using different parameters.

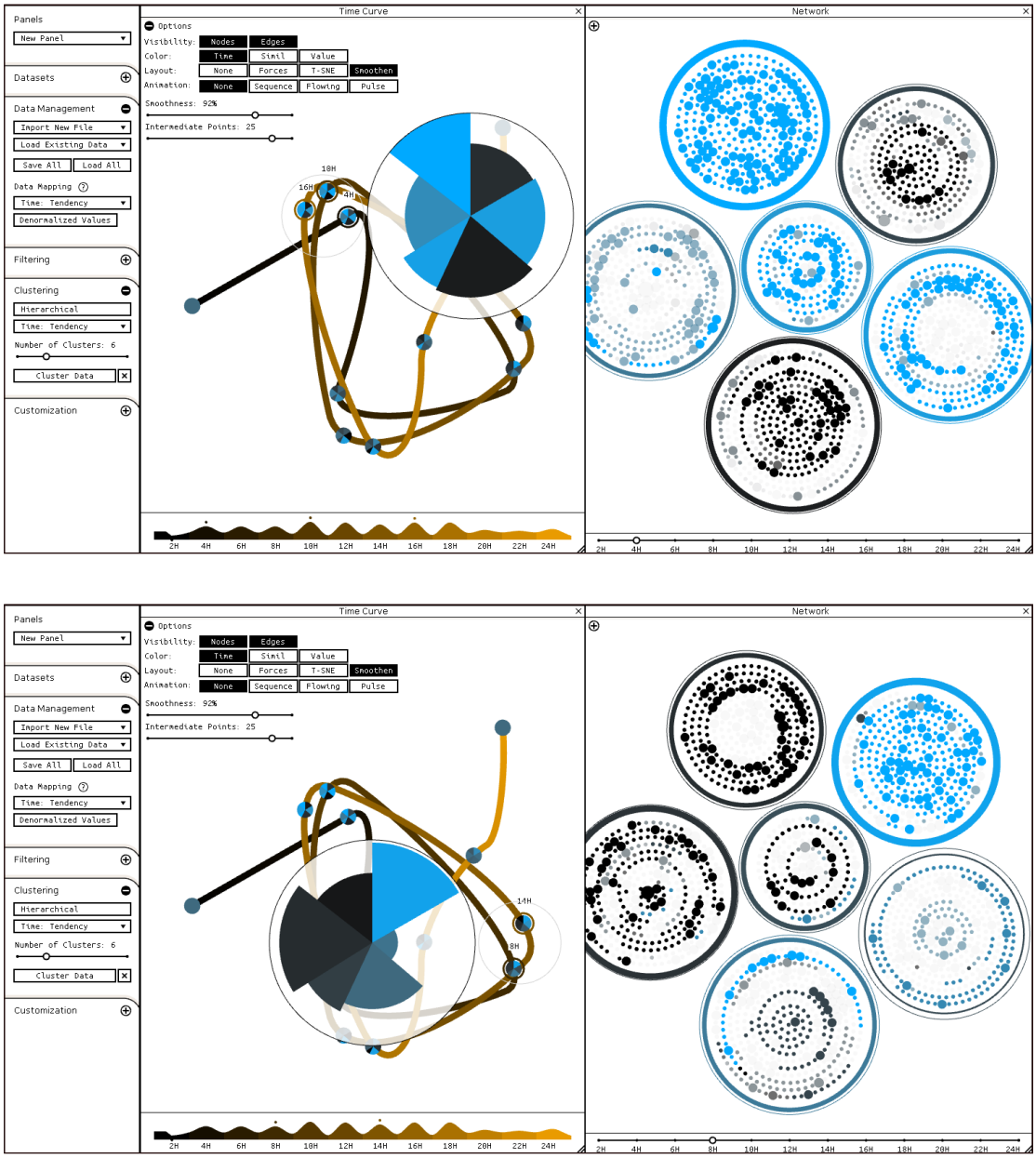
2.3.4. Timeline Graph

When any of the previous layout have been applied to the visualization, the slider at the bottom of the panel will transform into a graph. The waves represent the distance between sequential time points, where a large wave represents a time point that is very far from its previous one. Since distance equals dissimilarity, a large wave means that a significant change in the data happened at that moment. On the other hand, sections where the graph is very low mean that very few changes happened in the data, which would indicate a period of stability.

2.3.5. Cluster Glyphs

If the dataset has been clustered into multiple groups, the nodes in the visualization will be represented with pie chart glyphs. Each slice in the pie chart represents one cluster, and its color represents the average values of the data points in that

cluster at that time point. As such, each glyph is a simple representation of the state of the dataset for each time point. Glyphs with matching pie chart colors represent time points at which the dataset is behaving similarly. Additionally, if a cluster contains inactive nodes, the percentage of inactive nodes will be mapped to the transparency of the respective slice.



Time curve panel and network panel showing a clustered dataset. The mouse lens is hovered over two different groups of nodes (top and bottom). This creates a larger pie chart glyph and highlights similarity through slice radius and node trasparency in the network.

If glyphs are active, clicking the right mouse button on the time curve panel will create a circle around the mouse. This circle acts as a lens which selects every node inside of it and then displays a larger pie chart next to it that aggregates every selection.

This larger pie chart will display each slice with the average color of all the nodes selected. Additionally, the radius of the slice will also change based on the similarity of each cluster among the selected time points. This means that a slice with a large radius indicates that the data points belonging to a cluster present similar values across every selected data point. The opposite is also true, where very short slices represent significant shifts in data in that cluster between the selected times points.

This is further illustrated in the network panel. When multiple time points are selected with the lens, the transparency of individual nodes will match their similarity of values between the selected time points, highlighting those that behave similarly. Additionally, there will also be an outline with a variable weight, which matches the length of the radius of their respective slice on the pie charts.

The size of the lens can be increased or decreased with the mouse wheel, and the lens can be turned off by right-clicking a second time. Holding the CTRL key will keep every node selected, even when no longer inside the lens.

2.4. Variables Visualization

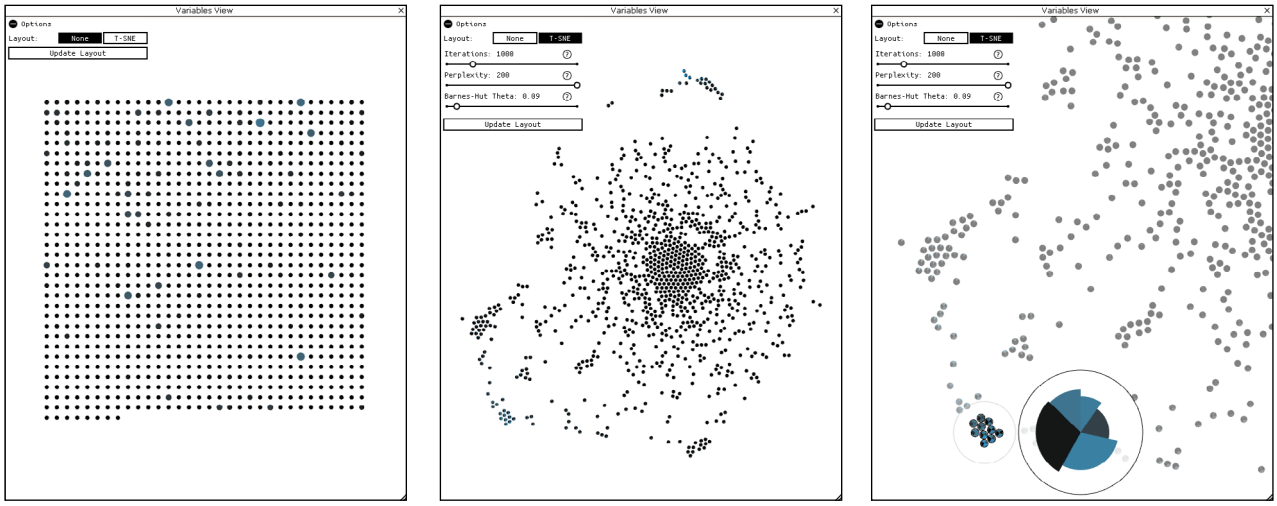
The variables panel will display every variable as a node in two-dimensional space. These variables can be any type of unordered propriety. The functionalities of this panel are similar to the time curve panel, where a layout is applied to position nodes relatively to their similarity, so that it may be possible to identify relationships between different variables.

2.4.1. Navigation

Variables are initially displayed in a grid layout, which helps distinguish the panel from others. Their initial color and size are mapped to the average value of that variable across the dataset. The mouse can also be dragged to pan the visualization

and the mouse wheel will zoom in/out of the current mouse position.

Clicking a node will select that variable in network panels, mapping every network node to the color and size of their respective value. Likewise, selecting a data point in either a data table or network panel will map the color and size of all the nodes in the variables panel to the variables of that data point.



The initial layout of the variables view panel displays each variable as a node in a grid (left), which can then be sorted by similarity of values with the T-SNE layout (middle). The mouse lens is also available in this panel, showing differences in variables between various clusters (right).

2.4.2. Variable Analysis

The top left of the panel contains an options panel where the layout of the nodes can be changed. The t-SNE layout will position nodes based on their values, placing nodes together if they have similar value patterns across different data points. This layout it is not dynamical and will only be calculated when the “Update” button is pressed. Additionally, it is also not deterministic, meaning that each recalculation of the layout will yield different results. The quality of these results can be influenced by the parameters, which balance accuracy and processing time. While each of these parameters is explained by hovering over the (?) icon, they are also addressed in Section 2.1.2.

If the dataset has been clustered into multiple groups, the nodes in the visualization will be represented with pie chart glyphs. Each slice in the pie chart represents one cluster, and its color represents the average values of the data points in that cluster at that time point. As such, each glyph is a simple representation of the state of the dataset for each variable, allowing them to be more easily compared. These glyphs are presented and explained in Section 2.3.5., where they are applied to time points instead of variables. Using right-click will create a circle around the mouse which acts as a lens for these glyphs, which is also addressed in Section 2.3.5.

3. Comparing Multiple Datasets

In CroP, multiple datasets can be stored individually and then visualized simultaneously. Through multiple visualization panels showing different datasets, data can be compared and potential patterns can be further analyzed. Additionally, the data table also allows the comparison between datasets with the same node names.

3.1. Adding Datasets

a new dataset using the “New Dataset” button under the “Data Management” section in the options sidebar. Existing datasets are shown as tabs which can be selected with a mouse click or scrolled through by dragging the mouse over them, if the number of tabs exceeds the size of the sidebar.

Options selected in the sidebar will only affect the currently selected dataset, including loading data. As such, separate data files can be uploaded into different dataset tabs. To change the dataset that is being displayed on a visualization panel, select the desired dataset from the dropdown located on the left of the top bar of every panel.

A selected portion of data from a loaded dataset can also be transferred onto a new tab, which can be used to visualize and analyze a section of the dataset considered significant without affecting previous work. This is achieved by selecting the relevant data points and then using the “Copy to New Dataset” button under “Filtering”. This will create a new dataset tab with a copy of the selected data.

3.2. Differences View

When multiple datasets have been loaded, a “Datasets” button will be added to data table panels at the top of the panel under the “General” tab. This button will open a table containing a list of every dataset loaded. Holding the CTRL key and selecting more than one dataset will activate a differences view in data table panels and network panels, which compares data points with the same names across all the selected datasets.

In data table panels, a tab will be added called “Datasets”, which contains a list of every common data point shared between all the selected datasets. One of the columns will contain a color matrix visualization of the time-series or variable data from each data point, but this one will be colored green by default. This color matrix represents the difference between the values across all selected datasets, where green represents fewer differences, and black represents higher differences. This column can be selected in order to order all the data points by how different they are. There is also an “Info” sub-tab containing information about the selected datasets, including an “Average Differences” visualization which depicts the average of all the differences across every common point.

In network panels, the color of every node will also be mapped to their differences across every dataset, just as in the data table. However, while the data table’s dataset tab does not show data points that aren’t common between all datasets, the network will still show those that belong to the currently selected dataset, although they will be transparent.

Contact Information

Developer: António Cruz *(antonioc@dei.uc.pt)*
Website: *https://cdv.dei.uc.pt/crop/*