FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Diplomarbeitspräsentation

# Guided Data Cleansing of Large Connectivity Matrices

Masterstudium:

Medical Informatics

Florence Gutekunst

Technische Universität Wien
Institut für Computergraphik und Algorithmen
Arbeitsbereich: Computergraphik
Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller

## Problem Statement / Motivation

The connectivity of the brain is stored in large connectivity matrices, which

- need to be **mined** to better understand how the brain works
- are **too large** to hold in current machine's memory
- contain **noisy and redundant data** → need to be **cleansed**

A visual tool is required for the user not to operate blindly on the choice of cleansing operation parameters. This cleansing is a step in the **connectivity matrices preprocessing**.

## Contribution

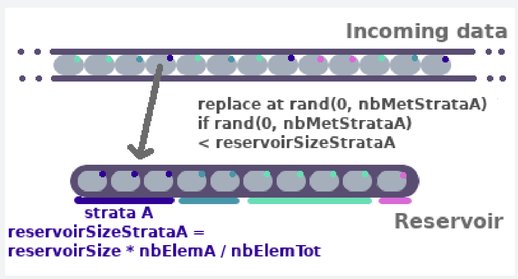
A visual tool was developed to **cleanse connectivity matrices**. It allows the user to

- define a threshold to eliminate noise
- define a similarity threshold to fuse together similar rows and columns
- have an overview of his data

## Sample the Connectivity Matrix

The connectivity matrix is too large to be stored in standard computer's memory. A **representative sampling** of the connectivity matrix needs to be performed first.

The rows and columns of the connectivity matrix represent brain neurons or groups of neurons. The value at each matrix cell represents the connectivity between the neurons represented by the row and column. The brain has a typical structure: it is **hierarchically built** [Spo16]. In order to get a representative sample, the sampling is performed **based on the anatomical hierarchy** of the rows and columns. The algorithm uses a **stratified reservoir sampling**, i.e., a subset of each group is randomly sampled in the reservoir, proportionally to the group size.

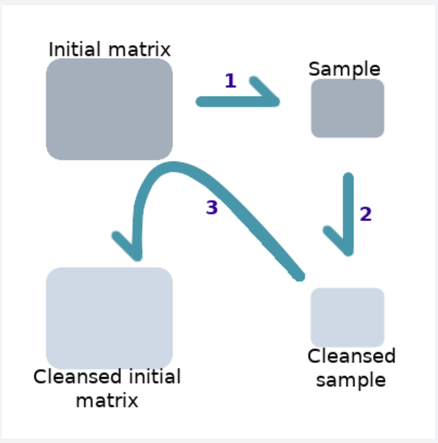



Stratified reservoir sampling

The sampling works as following:

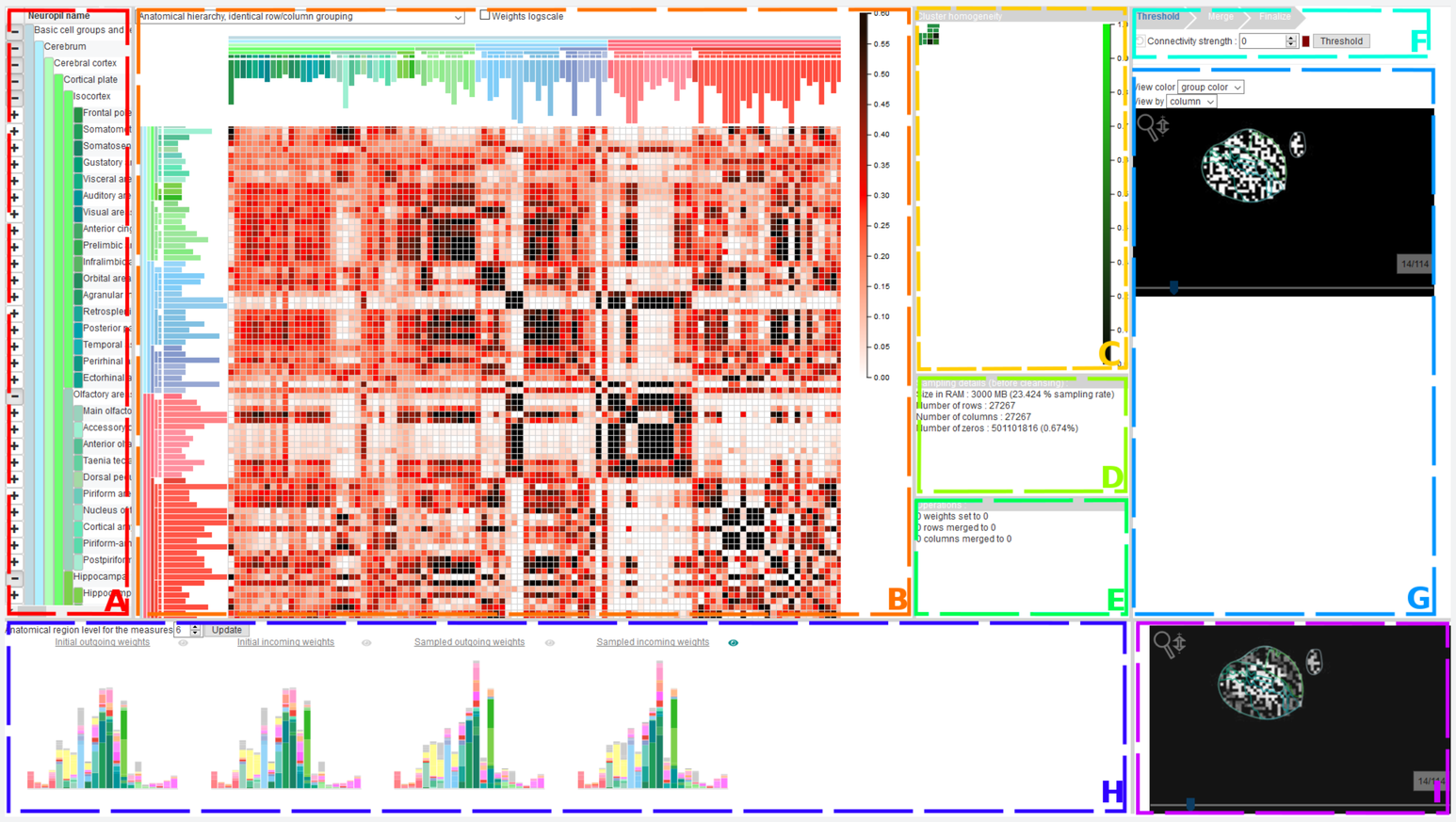
- the user defines an **anatomical hierarchy level**
- all anatomical structures on the given level are listed and the number of rows and columns to keep per anatomical structure are defined
- the rows and columns are **sampled according to their anatomical structure**

## Concept




1. The matrix is **sampled**
2. The sample is displayed and the user defines **cleansing parameters**
3. The cleansing parameters defined on the sample are **applied to the initial matrix**

## Visualize the Sample and Define Cleansing Parameters




- A Anatomical hierarchy view
- B Aggregated sample matrix view
- C Row / column cluster similarity view
- D Sampling details
- E Cleansing operations details
- F Cleansing operation control panel
- G Cleansing operation effects and spatial view of the sample
- H Network measures on the sample and cleansed sample
- I Spatial view of the sample colored according to the network measure results

The aggregated sample matrix view is inspired from MultiLayerMatrix [DCF16].

## Results

The three operations of the concept were evaluated on a mouse brain gene connectivity matrix (sampling, defining cleansing parameters, cleansing the initial matrix). **Network measures** were computed on all types of matrices (initial, sample, cleansed sample, cleansed initial) using the PAGANI Toolbox [DXZ+18]. The **distributions** of the network measures were then compared. A typical user (computer scientist in neuroscience) tested the tool and found it very easy to use.

The results show that at least one sixth of the initial matrix should be kept in the sample, and that the chosen anatomical hierarchical level should be **as deep as possible**.

## References

- T. N. Dang, H. Cui, and A.G. Forbes. Multilayermatrix: Visualizing large taxonomic datasets. *Proceedings of the 7th EuroVis Workshop on Visual Analytics*, 2016.
- H. Du, M. Xia, K. Zhao, X. Liao, H. Yang, Y. Wang, and Y. He. Pagani toolkit: Parallel graph-theoretical analysis package for brain network big data. *Human brain mapping*, 39(5):1869–1885, 2018.
- O. Sporns. *Connectome networks: from cells to systems*. Springer, 2016.

Kontakt: fgutekunst@free.fr