# Extracting the Groupwise Core Structural Connectivity Network: Bridging Statistical and Graph-Theoretical Approaches

**Abstract.** Finding the common structural brain connectivity network for a given population is an open problem, crucial for current neuroscience. Recent evidence suggests there's a tightly connected network shared between humans. Obtaining this graph will, among many advantages, allow us to focus cognitive and clinical analyses on common connections, thus increasing their statistical power. In turn, knowledge about the common graph will facilitate novel analyses to understand the structure-function relationship in the brain.

In this work we present a new algorithm for computing the core structural connectivity network of a subject sample combining graph theory and statistics. Our algorithm works in accordance with the novel evidence about brain topology. We analyze the problem theoretically and prove its complexity. Using 309 subjects, we show its advantages when used as a feature selection for connectivity analysis on populations, outperforming the current approaches.

**Keywords:** Group-wise connectome, core graph problem, brain connectivity, diffusion MRI

## 1 Introduction

Isolating the common brain connectivity network from a population is a crucial problem in current neuroscience [3, 7, 11]. Recent evidence suggests that there's a common and densely connected brain connectome across humans [2]. In this work we present a new approach for selecting this common connections, combining the recent knowledge of the problem and the current methods available [7, 11].

Finding the common brain connectome has the potential to increase our understanding of the relation between function and structure in the brain, one of the main questions in neuroscience [8]. Moreover, knowledge about the most common connections in a population will facilitate clinical and cognitive studies based on Diffusion MRI by reducing the number of connections to analyze. This reduction increases the statistical power of the studies. Finding the common connectome will also allow us to increase our knowledge about the brain structure by comparing core networks of different populations.

We formalize the problem of selecting the common connections using graph theory and statistics. Then we prove its NP-Hardness and propose a polynomial-time algorithm which finds an approximate solution. To do this, we develop an exact polynomial-time algorithm for a relaxed version of the problem and prove the algorithm's correctness and complexity. Then we adapt it to solve the general problem.

Currently, the most used algorithm for extracting a population's core structural connectivity network (CSNC) [7] uses an statistical approach: compute a connectivity matrix for each subject and analize each connection separately with a hypothesis test, using as null hypothesis that that edge is not present in the population. Finally, construct a binary graph with the edges for which the null hypothesis was rejected. The main problem of Gong et al.'s algorithm is that the resulting graph can be a set of disconnected subgraphs. But recent studies have shown that the brain has a *core* network tightly connected and a sparsely connected *outer* one [2]. In other words, this approach ignores the resulting network's topology. Moreover, doing statistics in a feature set chosen by hypothesis testing incurs in the double dipping problem [9].

A newer approach to solve the CSNC problem, designed by Wassermann et al. [11], uses graph theory to get a connected CSCN: compute a binary connectivity graph for each subject using a threshold, and for each possible connection compute the "cost" of including or excluding it from the common graph by evaluating in how many subjects that connection is present. Finally, construct the binary graph with all the edges that is "cheaper" to include than to exclude and connect the resulting graph if it's disconnected, using the minimum possible cost. This algorithm guarantees that the resulting graph is connected, but the connection binarization discards significant information for the resulting common network. In other words, it ignores the information of the probability of each connection being in the brain. This is problematic because the resulting graph may include edges for which the tractography has assigned a very low existence probability. Also, the outer part of the brain (the connections which do not end in the core network) should also be sparsely connected [2], which this algorithm does not enforce.

In this work we propose, for the first time, a polynomial algorithm to obtain the CSCN of a population which addresses the issues listed above. Our algorithm combines the recent graph-theoretical approach [11] with the statistical awareness of the most popular one [7]. We start by formalizing the problem, which allow us to prove that it's NP-Hard. Then, we propose a first algorithm that solves a relaxed version of the problem in an exact way, giving the best possible core graph for our formalization. Then, we adapt it to guarantee a connected result, agreeing with recent evidence on structural connectivity network topology (e.g [2]). Finally, we validate our approach using 300 subjects from the HCP database and comparing the performance of the networks obtained by our new approach, Wassermann et al.'s and Gong et al.'s predicting connectivity values from handedness in the core network.

## 2 Definitions, Problems and Contributions

As the problem we try to solve implies working with different brains, the first thing we need to do is unify them in a common connectivity model. This allows us to model all brains with graphs in which each node represents a cortical or sub-cortical region, and each edge represents a white matter connection between two regions. We chose the Desikan parcellation [4] to uniformize the brain cortical and sub-cortical regions across subjects.

We then use a probabilistic tractography algorithm to compute the connectivity matrix for each subject. This matrix represents, for each subject and brain connection, the existence probability of that connection [5]. Now we can model the entire sample networks as a set of weighted graphs with the same nodes.

Formally, a sample of $N$ brain structural networks can be represented by $N$ complete weighted graphs $G_1 = (V, E, w_1), \ldots, G_N = (V, E, w_N)$ with a common node set, which we call the *sample graphs*. Each graph $G_i$ corresponds to a subject. Each vertex $V$ represent a cortical or sub-cortical region. Each edge $E$ represent a white matter bundle connecting two regions ($E = V \times V$). And the weight $w_i(e)$ is the connection probability $e$ in the subject $i$ obtained through tractograpy ($w_1(e), w_2(e), \ldots, w_N(e) \in [0, 1]$ $\forall e \in E$). Note that all graphs have the same ordered node set and all of them are complete: the tractography output allows us to use the weight 0 to represent an absent connection.

Using this formalization we can express the general core structural connectivity network problem as follows: find a core graph $G^* = (V^*, E^*)$ densely connected such that $G^*$ keeps the more "relevant" connections $E^*$ in the sample and leaves out the less "relevant" ones ($E^* \subseteq E$), for some definition of relevance and some notion of density.

We want a formalization of relevance that abstracts the probability that a connection is present across subjects. Thus, we chose to model the relevance as the mean existence probability across subjects, factored by the standard deviation of these probabilities. In other words, $w^*(e)$ is the number of standard deviations that the mean of the connection $e$ is away from zero, which is a statistical measure of how present that edge is across the population.

$$w^*(e) = \frac{\overline{w(e)}}{s(e)} \tag{1}$$

where $\overline{w(e)}$ is the population mean of weights of that connection across subjects and $s(e)$ the sample standard deviation of that connection across subjects.

Note that $w^*$ (eq. 1) is the statistic of a hypothesis z-test which assumes a media of 0 for the weights. We chose the z-statistic because of the normal distribution's properties, e.g. linearity, even if a Beta or Gamma distribution may be more appropriate for modeling the probability. In any case, note that for the purpose of our contribution $w^*$ can be any function $V \times V \to \mathbb{R}$ which grows with the relevance of the edges in the sample.

To represent the density of the core subgraph we used the relationship between the number of edges and the total statistical relevance $w^*$ that those edges sum:

$$\alpha(w^*, E^*) = \frac{\sum_{e \in E^*} w^*(e)}{|E^*|} \tag{2}$$

As we want also a sparse outer subgraph, we also define its density:

$$\beta(w^*, E^*) = \frac{\sum_{e \in E \setminus E^*} w^*(e)}{|E^*|} \tag{3}$$

Now we can express our objective informally as: choose $E^*$ such that $\alpha(w^*, E^*)$ (eq. 2) is large and $\beta(w^*, E^*)$ (eq. 3) small. Note that once we chose $E^*$ we can define $V^*$ as all the vertices that have some edge in $E^*$. This simplifies the problem of finding $G^*$ to find $E^*$ alone. We also want $G^*$ to be connected. Here, connected means that for every pair of vertices $u, v$ in $V^*$ there is a path of edges in $E^*$ from $u$ to $v$.

We can formalize the goal of finding this common graph $G^*$ in two different ways.

- Optimization version:
$$\max_{E^* \subseteq E} f(w^*, E^*) = \lambda\alpha(w^*, E^*) - (1-\lambda)\beta(w^*, E^*) \tag{4}$$
restricting the solution to connected graphs. Here, $\lambda$ is a parameter between 0 and 1 which can be adjusted to weight the density of the inner and the outer network. Note that if $\lambda = 1$, the solution to (4) only considers the density of the core network, and if $\lambda = 0$, it only considers the edges excluded of the core network.
- Decision version: Given $A, B$ find $E^*$ connected such that:
$$\begin{aligned} \alpha(w^*, E^*) \geq A \\ \beta(w^*, E^*) \leq B \end{aligned} \tag{5}$$

Now we will prove that the problem is NP-Complete.

### 2.1 NP-Completeness Result

We have formalized the problem of the Core Structural Connectivity Network taking into account the density and connectedness of the core subgraph and the sparsity of the outer one. We will now prove that, with this formalization, the problem is NP-Complete.

**Definition 1 (Core Structural Connectivity Network problem).** *Given $G_1 = (V, E, w_1), G_2 = (V, E, w_2), \ldots, G_N = (V, E, w_N)$ weighted graphs (the* sample graphs*) with a common node set, a complete edges set ($E = V \times V$) and $w_1(e), w_2(e), \ldots, w_N(e) \in \mathbb{R}_{\geq 0}$ $\forall e \in E$ weights of their edges, and given $A, B$ real numbers, find $G^* = (V^*, E^*)$ connected graph (the* core graph*) such that*

$$\alpha(w^*, E^*) \geq A$$

$$\beta(w^*, E^*) \leq B$$

*for $\alpha$ and $\beta$ as defined in (2) and (3).*

Here we prove that the *Core Structural Connectivity Network problem*, called CSCN problem, is NP-complete. In our reduction, we use the *Steiner Tree problem* [6], called ST problem in the following. Given an edge-weighted graph

$G' = (V', E', w)$, a subset $S \subseteq V'$ of nodes, and a real $k \geq 0$, ST problem consists in determining if there exists a connected subgraph $H$ such that $S \subseteq V(H)$ and $\sum_{e \in E(H)} w(e) \leq k$. The decision version of ST problem is NP-complete even if all weights are equal [6].

**Instance of ST problem.** Consider any edge-weighted graph $G' = (V', E', w)$ such that $w(e) = \frac{1}{2}$ for every $e \in E'$. Given $k \geq 0$, ST problem consists in determining if there exists a connected subgraph $H$ such that $S \subseteq V(H)$ and $|E(S)| \leq 2k$. Without loss of generality, we assume that $|E'| \geq 2k$ and that $G'$ is connected.

**Reduction.** We construct the instance of CSCN problem as follows. Let $s = |S|$ and let $t \geq 1$ be any positive integer. Let $G = (V, E, w^*)$ defined as follows. Let $V = V' \cup \{v_{i,j} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$ and $E = V \times V$. Let $S = \{u_1, \ldots, u_s\}$. For every $i, j$, $1 \leq i \leq s$, $1 \leq j \leq t$, $w^*_{v_{i,j},u_i} = 1$ and $w^*_{v_{i,j},u} = 0$ for every $u \in V \setminus \{u_i\}$. Furthermore, for every $e \in E'$, set $w^*(e) = w(e) = \frac{1}{2}$, and for every $u, u' \in V'$ such that $\{u, u'\} \notin E'$, then set $w^*_e = 0$. Finally, we set $A = \frac{s.t+k}{s.t+2k}$ and $B = \frac{\frac{1}{2}(|E'|-2k)}{s.t+2k}$.

**Lemma 1.** *If $|E^*| < s.t + 2k$, then any solution for CSCN problem is not admissible because $\beta(w^*, E^*) > B$.*

**Proof**. Suppose that $|E^*| < s.t + 2k$. In order to minimize $\sum_{e \in E \setminus E^*} w^*(e)$, $E^*$ must contain $\{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$ if $|E^*| \geq s.t$. (Otherwise we select a subset of this set of edges.) Indeed, by construction of $G$, we have $w^*_{v_{i,j},u_i} = 1$ for every $i, j$, $1 \leq i \leq s$, $1 \leq j \leq t$. Then, if $|E^*| - s.t > 0$, $E^*$ must contain $|E^*| - s.t$ edges of $E'$, that is edges of $E$ of weight $\frac{1}{2}$ each. Recall that there are exactly $s.t$ edges of weight 1, and the other edges have weight 0 or $\frac{1}{2}$.

There are two cases. First, suppose that $|E^*| \geq s.t$. We get that $\sum_{e \in E \setminus E^*} w^*(e) = \frac{1}{2}(|E'| - (|E^*| - s.t))$. Since $|E^*| - s.t < 2k$, then we get that $\sum_{e \in E \setminus E^*} w^*(e) = \frac{1}{2}(|E'| - (|E^*| - s.t)) > \frac{1}{2}(|E'| - 2k)$. Furthermore, since $|E^*| < s.t + 2k$, we get that $\frac{\frac{1}{2}(|E'|-(|E^*|-s.t))}{|E^*|} > \frac{\frac{1}{2}(|E'|-2k)}{s.t+2k}$. Thus, we proved that $\beta(w^*, E^*) = \frac{\sum_{e \in E \setminus E^*} w^*(e)}{|E^*|} > B$.

Second, suppose that $|E^*| < s.t$. We get that $\sum_{e \in E \setminus E^*} w^*(e) = s.t - |E^*| + \frac{|E'|}{2}$. Since $s.t - |E^*| + \frac{|E'|}{2} > \frac{1}{2}(|E'| - (|E^*| - s.t))$, we obtain the result by the arguments described for the first case.

Finally, if $|E^*| < s.t + 2k$, then there is no admissible solution for CSCN problem. □

**Lemma 2.** *If $|E^*| > s.t + 2k$, then any solution for CSCN problem is not admissible because $\alpha(w^*, E^*) < A$.*

**Proof**. Suppose that $|E^*| > s.t + 2k$. In order to maximize $\sum_{e \in E^*} w^*(e)$, $E^*$ must contain $\{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$ and $|E^*| - s.t$ edges of $E'$. Indeed, by construction of $G$, we have $w^*_{v_{i,j},u_i} = 1$ for every $i, j$, $1 \leq i \leq s$, $1 \leq j \leq t$. Furthermore, $w^*(e) = \frac{1}{2}$ for every $e \in E'$. Recall that there are exactly $s.t$ edges of weight 1, and the other edges have weight 0 or $\frac{1}{2}$. We get that $\alpha(w^*, E^*) = \frac{s.t+\frac{1}{2}(|E^*|-s.t)}{|E^*|} < \frac{s.t+k}{s.t+2k} = A$. Indeed, the average weight is lower when there are more edges of weight $\frac{1}{2}$ (the number of edges of weight 1 is the same in both ratios).

Finally, if $|E^*| > s.t + 2k$, then there is no admissible solution for CSCN problem. □

By Lemma 1 and Lemma 2, we get the following corollary.

**Corollary 1.** *Any solution for CSCN problem is such that* $|E^*| = s.t + 2k$.

We prove in Lemma 3 and in Lemma 4 that there is an admissible solution for CSCN problem if and only if there is an admissible solution for ST problem.

**Lemma 3.** *If there is an admissible solution for ST problem, then there is an admissible solution for CSCN problem.*

**Proof**. Suppose there is an admissible solution for ST problem. We prove that there is an admissible solution for CSCN problem. Let $H$ be a connected subgraph such that $S \subseteq V(H)$ and $\sum_{e\in E(H)} w(e) = \frac{1}{2}|E(|H)| \leq k$. If $|E(H)| = 2k$, then set $E^* = E(H) \cup \{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$. If $|E(H)| < 2k$, then set $E^* = E(H) \cup \{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\} \cup F$, where $F \subseteq E'$ such that $F \neq E(H) = \emptyset$, $w^*(e) = \frac{1}{2}$ for every $e \in F$, and such that the graph induced by $E(H) \cup F$ is connected. The last condition comes from Corollary 1 in order to get the right number of edges in $E^*$. This condition is always possible to satisfy because $G'$ is connected.

The graph induced by $E^*$ is connected. Indeed, $H$ is an admissible solution for ST problem, $E(H) \cup F$ is connected by construction, and $\{\{v_{i,j}, u_i\} \mid 1 \leq j \leq t\}$ is a set of edges all adjacent to $u_i \in S$ for every $i$, $1 \leq i \leq s$.

Furthermore, we get

$$\alpha(w^*, E^*) = \frac{\sum_{e\in E^*} w^*(e)}{|E^*|} = \frac{s.t+k}{s.t+2k} = A$$

and

$$\beta(w^*, E^*) = \frac{\sum_{e\in E\setminus E^*} w^*(e)}{|E^*|} = \frac{\frac{1}{2}(|E'|-2k)}{s.t+2k} = B.$$

Finally, we proved that there is an admissible solution for CSCN problem. □

**Lemma 4.** *If there is an admissible solution for CSCN problem, then there is an admissible solution for ST problem.*

**Proof**. Suppose there is an admissible solution for CSCN problem. We prove that there is an admissible solution for ST problem. Let $E^* \subseteq E$ be such that the graph induced by $E^*$ is connected, and such that $\alpha(w^*, E^*) \geq \frac{s.t+k}{s.t+2k} = A$ and $\beta(w^*, E^*) \leq= \frac{\frac{1}{2}(|E'|-2k)}{s.t+2k} = B$.

We first prove that $\{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\} \subseteq E^*$. By Corollary 1, we know that $|E^*| = s.t + 2k$. Thus, it necessarily means that $\sum_{e \in E^*} w^*(e) \geq s.t + k$. By construction of $G$, the set of edges of weight 1 is $\{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$, that is $\{e \in E \mid w^*(e) > \frac{1}{2}\} = \{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\}$. We get that $\{\{v_{i,j}, u_i\} \mid 1 \leq i \leq s, 1 \leq j \leq t\} \subseteq E^*$ because, otherwise, we would have $\sum_{e \in E^*} w^*(e) < s.t + k$.

Furthermore, every $e \in E^* \cap E'$ is such that $w^*(e) = \frac{1}{2}$. Indeed, otherwise, we would have $\sum_{e \in E^*} w^*(e) < s.t + k$.

Finally, since $E^*$ is an admissible solution for CSCN problem, then it means that the graph induced by the set of edges $E^* \cap E'$ is connected and is such that for every $u_i$, $1 \leq i \leq s$, then there is an edge $e \in E^* \cap E'$ that is adjacent to $u_i$. By the previous remark, every edge in $E^* \cap E'$ has weight $\frac{1}{2}$. Thus, it means that there is $E^* \cap E'$ is an admissible solution for ST problem considering the graph $G'$. Indeed $|E^* \cap E'| = 2k$ and so $\sum_{e \in E^* \cap E'} w(e) = k$. □

We are now able to prove the NP-completeness of CSCN problem.

**Theorem 1.** *CSCN problem is NP-complete.*

**Proof**. The reduction is clearly polynomial. Furthermore, Lemma 3 and Lemma 4 prove the equivalence between CSCN problem and ST problem. Since the decision version of ST problem is NP-complete even if all weights are equal [6], then we obtain the NP-completeness of the decision version of CSCN problem. □

As an illustration of the proof of Theorem 1, consider the instance of ST problem and CSCN problem depicted in Figure 1. In this example, we set $s = 5$, $t = 3$, $k = 3$. The edges of $E'$ have weight $\frac{1}{2}$, the edges of $E$ have weight 1, and the other edges (non-edges in Figure 1) have weight 0. We set $A = \frac{6}{7}$ and $B = \frac{1}{3}$. There is an admissible solution for CSCN problem with $|E^*| = 21$ and so an admissible solution for ST problem with $|E(H)| = 6$. We get $\alpha(w^*, E^*) = A$ and $\beta(w^*, E^*) = B$.

In Theorem 1 we have proved that CSCN problem is $\mathsf{NP}$-complete. Hence, to be able to solve it in reasonable time we need a relaxation to make it tractable or an approximate algorithm for the complete version. In this article we will propose both.

### 2.2 Relaxation approach

We found that the connectivity constraint is the main reason of the difficulty of the problem. Without it, it becomes tractable.

**Theorem 2.** *The decision version of CSCN problem without the connectivity constraint is in* $\mathsf{P}$.

**Proof**.

The algorithm 1, in each step $i$, defines $E^*$ as the $i$ maximum weighted edges and tries to use that to fulfill the constraints.

Assume that there exists an $E^*$ that fulfills the constraints. Two cases exist:

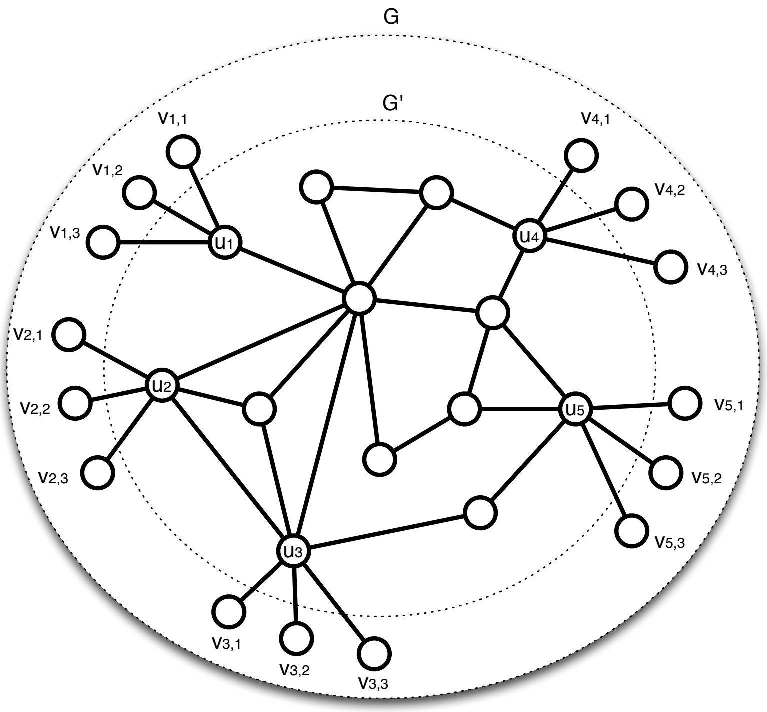

**Fig. 1.** Illustration of the proof of Theorem 1

1. $E^*$ has the $|E^*|$ maximum weighted edges.
2. There are $e_j \in E^*$, $e_k \in E \setminus E^*$ such that $w^*(e_k) \geq w^*(e_j)$.

In the first case, the algorithm 1 will find $E^*$.

In the second one, let $E' = (E^* \cup \{e_k\}) \setminus \{e_j\}$ another subset of $E$. Then

$$\alpha(w^*, E') = \frac{\sum_{e \in E'} w^*(e)}{|E'|} = \frac{\sum_{e \in E'} w^*(e)}{|E^*|} \geq \frac{\sum_{e \in E^*} w^*(e)}{|E^*|} = \alpha(w^*, E^*) \geq A$$

because the edges in $E^*$ are the same as the ones in $E'$ except from one that has a larger weight. For the same reason,

$$\beta(w^*, E') = \frac{\sum_{e \in E \setminus E'} w^*(e)}{|E'|} = \frac{\sum_{e \in E \setminus E'} w^*(e)}{|E^*|} \leq \frac{\sum_{e \in E \setminus E^*} w^*(e)}{|E^*|} = \beta(w^*, E^*) \leq B$$

Thus, we found a new subset of $E$ that stills fulfills the constraints. We can do the same process with $E'$ (replace an edge with another one of larger weight) iteratively, always getting subsets that fulfills the constraints, until we can't do

**Algorithm 1** Maximum edges

```
Compute $w^*(e)$ for each $e \in E$
Sort($E$)                                    ▷ sorts edges by $w^*$ non-increasingly
for each $e \in E$ do
    $E^* \leftarrow E^* \cup e$
    if $\alpha(w^*, E^*) > A$ and $\beta(w^*, E^*) < B$ then
        return $True$
    end if
end for
return $False$
```

this anymore. At that point we'll have a subset that has only the maximum $|E^*|$ edges and fulfills the constraints. Thus, algorithm 1 will find this subset.

We now need to prove algorithm 1 runs in polynomial time in the size of $|E|$. The first operation, computing $w^*(e)$ for each $e$, implies computing the mean and standard deviation for each edge across the population, which can be done in $\mathcal{O}(N)$ per edge (where $N$ is the size of the population). This is $\mathcal{O}(N * |E|)$ for all the edges. The second step, sorting, can be done in $\mathcal{O}(|E| \log |E|)$.

The main loop runs at most $|E|$ times, and in each loop it adds an edge to $E^*$, computes $\alpha$ and $\beta$ and performs two comparisons. The comparisons can be done in constant time, as the addition to $E^*$ if we use a linked list of edges to represent it. To compute $\alpha$ and $\beta$ it is needed to iterate once again $E$ (the part in $E^*$ for $\alpha$, the part in $E \setminus E^*$ for $\beta$) adding the weights together and then performing two divisions. This can be done in linear time in the size of $E$, and even quicker (constant time) if we optimize it by keeping the values of $\alpha$ and $\beta$ across loops and updating them with the weight of the edge that changed sets.

Then, algorithm 1 solves the CSCN problem in $\mathcal{O}(max(|E|^2, |E| * N)$ or in $\mathcal{O}(|E| * N)$ if a little optimization is used.

□

### 2.3 Heuristic approach

We gave an algorithm that solves the problem of finding the Core Structural Connectivity Network in polynomial time but without guaranteeing a connected result. We can solve the complete problem by applying algorithm 1 and then connecting the resulting core graph, if it wasn't connected in the first place. This gives us an approximate solution for the problem in polynomial time.

For connecting the graph we want to add edges decreasing the objective function the minimum possible. For this, we use the same approach that Wassermann et al. [11]. Namely, we add the edges between connected components using a Maximum Spanning Tree algorithm using $w^*$ (eq. 1) as the weights of the edges. This way we get a connected subgraph close to the best possible subgraph, which we got using algorithm 1.

## 3 Experiments and Results

Now we will asses the performance of our method, compared with the most used [7] and with the recent one [11] in the task of connectivity prediction performance.

We used the HCP500 dataset [10]: 309 subjects aged 21-40 with complete dMRI protocol. We obtained the weighted connectivity matrices between the cortical regions defined by the Desikan atlas [4] as done by Basset et al. [1].

### 3.1 Predicting Handedness-specific Connectivity

**Table 1.** Amount of features selected by linear regression in the core graph, relating the weights with handedness. Our procedure gets more features selected than Gong et al. [7] and Wassermann et al. [11], showing better statistical power.

| Algorithm | Mean | Std |
|---|---|---|
| Gong et al. 2009 | 0.066 | 0.256339 |
| Wassermann et al. 2016 | 0.415 | 0.723394 |
| Our approach ($\lambda = 0.50$) | 1.042 | 1.268797 |

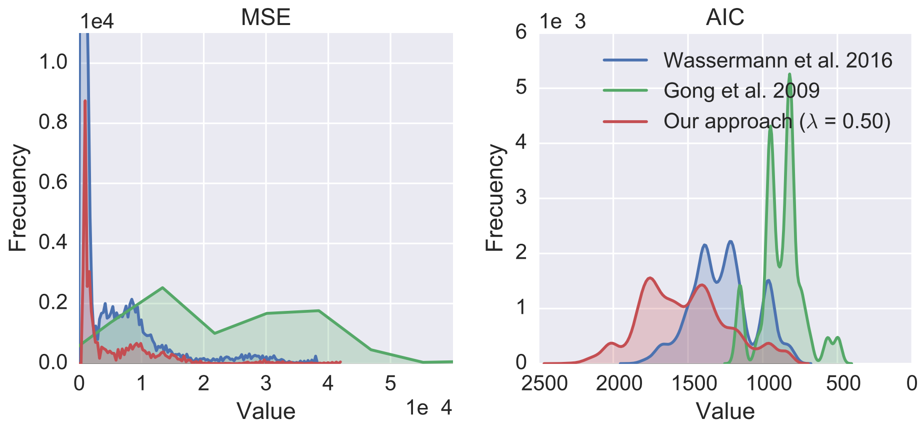

**Fig. 2.** Performance of core network as feature selection for a linear model for handedness specific connectivity. We evaluate model prediction (left) and fit (right) for Gong et al. [7] in green, Wassermann et al. [11] in blue and ours, in red. We show the histograms of both values from our nested Leave-$\frac{1}{3}$-Out experiment. In both measures, our approach has more frequent lower values than Gong et al., showing a better performance.

We evaluated the performance of our method using the generated core graphs as a feature selection for fitting a linear model with handedness specific connectivity. We used a nested Leave-$\frac{1}{3}$-Out procedure: the outer loop performs model selection on $\frac{1}{3}$ of the subjects using the core graph algorithm and the inner loop performs model fitting and prediction using the selected features.

More in detail, we first take $\frac{1}{3}$ subjects randomly and compute the core graph for those subjects using the three different algorithms. Then we add the weights for the selected edges for each subject, and select the features $F$ that are more determinant of handedness using a linear least-squares regression and the Bonferroni correction for multiple hypothesis. We quantify the amount of features that are selected after this procedure, which indicates how useful is the core graph algorithm for selecting the edges related to handedness.

We then randomly take $\frac{1}{2}$ of the remaining subjects and fit a linear model on $F$ for predicting connectivity weights using the handedness of each subject. Finally, we predict the values of the features $F$ from the handedness column in the subjects left and quantify the prediction performance with the mean squared error (MSE) of the prediction and Akaike Information Criterion (AIC) for model fitting. For both measures a lower value indicates better performance. The outer loop is performed 500 times and the inner loop 100 times per outer loop, which totals 50,000 experiments.

We show the experiments' results in Table 1 and in Fig. 2. In Table 1 we can see that our algorithm preserves connections correlated with the handedness of the subjects. This makes our approach a good feature selector for the task. In Fig. 2 we see that our method performed better than Gong et al. [7] as the number of cases with lower AIC and MSE is larger in our case. This means our method works better fitting a model with handedness and predicting edge connectivity related to it. In summary, our algorithm outperforms others when used as feature selection for this connectivity analysis.

## 4 Discussion and Conclusion

We presented for the first time a polynomial algorithm to extract the core structural connectivity network of a population combining a graph-theoretical approach with statistic relevance of the connections, observing the recent evidence of the structural network topology. We formalized the problem, proved it's difficulty and gave a novel algorithm for dealing with it. We then validated our approach by showing its power as feature selector for getting connections related to handedness with 300 real subjects' data. The experiment shows our method performs better than the currently available. We leave the door open to further research of sound validations for this type of algorithms.

## References

1. Bassett, D.S., Brown, J.A., Deshpande, V., Carlson, J.M., Grafton, S.T.: Conserved and variable architecture of human white matter connectivity. NeuroImage 54(2), 1262–1279 (2011), http://dx.doi.org/10.1016/j.neuroimage.2010.09.006

2. Bassett, D.S., Porter, M.A., Wymbs, N.F., Grafton, S.T., Carlson, J.M., Mucha, P.J.: Robust detection of dynamic community structure in networks. Chaos 23(1) (2013)
3. Bullmore, E.T., Sporns, O., Solla, S.A.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nature reviews. Neuroscience 10(3), 186–98 (2009), http://www.ncbi.nlm.nih.gov/pubmed/19190637
4. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J.: An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31(3), 968–980 (2006)
5. Donahue, C.J., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Behrens, T.E., Dyrby, T.B., Coalson, T., Kennedy, H., Knoblauch, K., Van Essen, D.C., Glasser, M.F.: Using Diffusion Tractography to Predict Cortical Connection Strength and Distance: A Quantitative Comparison with Tracers in the Monkey. Journal of Neuroscience 36(25), 6758–6770 (jun 2016), http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0493-16.2016
6. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1979)
7. Gong, G., He, Y., Concha, L., Lebel, C., Gross, D.W., Evans, A.C., Beaulieu, C.: Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging tractography. Cerebral Cortex 19(3), 524–536 (2009)
8. Knock, S.A., McIntosh, A.R., Sporns, O., Kötter, R., Hagmann, P., Jirsa, V.K.: The effects of physiologically plausible connectivity structure on local and global dynamics in large scale brain models. Journal of Neuroscience Methods 183(1), 86–94 (2009)
9. Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I.: Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience (2009)
10. Sotiropoulos, S.N., Jbabdi, S., Xu, J., Andersson, J.L., Moeller, S., Auerbach, E.J., Glasser, M.F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D.A., Yacoub, E., Lenglet, C., Van Essen, D.C., Ugurbil, K., Behrens, T.E.: Advances in diffusion MRI acquisition and processing in the Human Connectome Project. NeuroImage 80(3), 125–143 (oct 2013), http://linkinghub.elsevier.com/retrieve/pii/S105381191300551X
11. Wassermann, D., Makris, N., Rathi, Y., Shenton, M., Kikinis, R., Kubicki, M., Westin, C.F.: The white matter query language: a novel approach for describing human white matter anatomy. Brain Structure and Function pp. 1–17 (2016)