HAL open science

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

Nahuel Lascano, Guillermo Gallardo, Rachid Deriche, Dorian Mazauric, Demian Wassermann

To cite this version:

Nahuel Lascano, Guillermo Gallardo, Rachid Deriche, Dorian Mazauric, Demian Wassermann. Extracting the Groupwise Core Structural Connectivity Network: Bridging Statistical and Graph-Theoretical Approaches. 2017. hal-01426870v2

HAL Id: hal-01426870

https://inria.hal.science/hal-01426870v2

Preprint submitted on 5 Jan 2017 (v2), last revised 7 Jan 2017 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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

Nahuel Lascano[1,2], Guillermo Gallardo[1], Rachid Deriche[1], Dorian Mazauric[3], and Demian Wassermann[1]

[1] Athena EPI, Université Côte d'Azur, Inria, France
[2] Computer Science Department, FCEyN, Universidad de Buenos Aires, Argentina
[3] ABS EPI, Université Côte d'Azur, Inria, France

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 network 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 network 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 novel evidence on 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 main problem in current neuroscience [3, 7, 10]. 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 these common connections, combining recent topological hypotheses [2] and current methods [7, 10].

Finding the common brain connectome across subjects has the potential to increase our understanding of the relationship between function and structure in the brain. This relationship is one of the main open questions in neuroscience [3, 5]. Moreover, knowledge about the most common connections in a population will facilitate clinical and cognitive Diffusion MRI analyses by reducing the number of surveyed connections, increasing the statistical power of those analyses. Finding

the common connectome will also allow us to increase our knowledge about the brain structure by comparing core networks across different populations.

We formalize the problem of selecting the common connections combining graph theory and statistics. Then, we prove that the problem is NP-Hard and propose a polynomial-time algorithm to find approximate solutions. 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.

Currently, the most used algorithm to extract a population's core structural connectivity network (CSNC) [7] uses an statistical approach: first, compute a connectivity matrix for each subject; then, 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 [7] algorithm is that the resulting graph can be a set of disconnected subgraphs. Moreover, 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. Performing statistical analyses in a feature set chosen by hypothesis testing incurs in the double dipping problem [8].

A newer approach to solve the CSNC problem, designed by Wassermann et al. [10], uses graph theory to get a connected CSCN: first, compute a binary connectivity graph for each subject using a threshold; 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 discards information of the probability of each connection being in the brain. This is problematic because the resulting graph may include edges for which tractography assigned a very low existence probability across subjects. Also, the outer part of the brain, the connections which do not result 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-time algorithm to obtain the CSCN of a population addressing the issues listed above. Our algorithm combines the recent graph-theoretical approach [10] 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 [2, e.g.]. 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 [10] and Gong et al.'s [7] predicting connectivity values from handedness in the core network.

2 Definitions, Problems and Contributions

We want to develop a new algorithm to extract the core structural connectivity network, a problem that implies working with different brains. Thus, the first thing we need to do is unify them into 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 choose the Desikan parcellation [4] to uniformize the brain cortical and sub-cortical regions across subjects.

To compute the connectivity matrices we use a probabilistic tractography algorithm, which outputs one matrix per subject. The resulting matrices represent the existence probability of a connection across parcels in each subject [5]. As these are symmetric, we interpret the matrices as weighted undirected graphs, sharing the node set across subjects.

Formally, we represent a sample of N brain structural networks by N complete weighted graphs $G_1 = (V, E, w_1), \ldots, G_N = (V, E, w_N)$ with a common node set V. We call $G_1, \ldots, G_N$ the *sample graphs*. Each graph G_i corresponds to a subject. Each vertex $v \in V$ represents a cortical or sub-cortical region. Each edge $e \in E = V \times V$ represents a white matter bundle connecting two regions. Finally, the weight $w_i(e)$ is the connection probability for the edge e in the subject i obtained through tractography:

$$w_1(e), w_2(e), \ldots, w_N(e) \in [0, 1] \ \forall e \in E. \tag{1}$$

Note that all graphs have the same ordered node set and all of them are complete: an edge weight, or connection probability, $w_i(e)$ of 0 represents an absent connection. Using this formalization we 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 $E^* \subseteq E$ in the sample and discards the less *relevant* ones, for some definition of relevance and density. For simplicity, once we select E^* we can define V^* as

$$V^* = \{v \in V : \exists u \in V, (u, v) \in E^*\} \ , \tag{2}$$

the set of nodes that the edges in E^* cover. Then, we can reduce the problem of finding G^* to find E^* alone.

We want a formalization of relevance that represents the probability that a connection is present across subjects. Thus, we choose to model the groupwise relevance, $w^*(e)$ 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 existing probability of connection e is larger than 0. We use $w^*(e)$ as a statistical measure of edge presence across the population. Formally,

$$w^*(e) \triangleq \frac{\overline{w(e)}}{s(e)} \text{ where } \overline{w(e)} \triangleq \sum_{i=1}^{N} \frac{w_i(e)}{N}, \ s(e) \triangleq \sqrt{\sum_{i=1}^{N} \frac{\left(\overline{w(e)} - w_i(e)\right)^2}{N}}. \tag{3}$$

Note that $w^*(e)$ is the statistic of a hypothesis z-test which assumes a media of 0 for the population weight of e. We choose the z-statistic because of the normal distribution's properties, e.g. linearity, even if other distributions, such as Beta 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 $E \to \mathbb{R}$ which grows with the relevance of the edges in the sample.

We also want a formalization that represents the density of the core subgraph. We use the relationship between the number of edges and the total statistical relevance w^* that those edges sum:

$$\alpha(w^*, E^*) \triangleq \frac{\sum_{e \in E^*} w^*(e)}{|E^*|} \ . \tag{4}$$

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

$$\beta(w^*, E^*) \triangleq \frac{\sum_{e \in E \setminus E^*} w^*(e)}{|E^*|} \ . \tag{5}$$

Now we can express our objective informally as: choose E^* such that $\alpha(w^*, E^*)$ (Eq. 4) is large and $\beta(w^*, E^*)$ (Eq. 5) small. In accordance to recent evidence on the core network, 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.

Let $\mathcal{E}^c$ be the family of sets of edges that induce a connected graph. We now formalize the problem of finding this common graph G^* in two different ways.

- The optimization version consists in computing:

$$\max_{E^* \in \mathcal{E}^c} f(w^*, E^*) = \lambda \alpha(w^*, E^*) - (1 - \lambda)\beta(w^*, E^*) \tag{6}$$

  The parameter λ (between 0 and 1) can be adjusted to weight the density of the inner and the outer network. Note that if $\lambda = 1$, the solution to (6) only considers the density of the core network, and if $\lambda = 0$, it only considers the edges excluded of the core network.
- Given A and B, the decision version consists in finding $E^* \subseteq \mathcal{E}^c$ such that:

$$\begin{aligned} \alpha(w^*, E^*) &\geq A \\ \beta(w^*, E^*) &\leq B \end{aligned} \tag{7}$$

Having formalized the Core Structural Connectivity Network into an optimization and a decision problem, we proceed with one of our main theoretical contributions: proving that the problem is NP-Complete.

2.1 CSCN Problem's NP-Completeness

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 α *and* β *as defined in Eq. (4) and Eq .(5).*

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. □

In Theorem 1 we have proved that CSCN problem is 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 of the CSCN problem

We proved in the previous section that the connectivity constraint is the main reason of the difficulty of the problem. Without it, it becomes tractable. So we solve, in this section, a relaxed version of problem without the connectivity constraint. Then, we use this solution to approximate the full problem.

Theorem 2. *The decision version of CSCN problem without the connectivity constraint is in* P.

Algorithm 1 Maximum edges

```
Compute w*(e) for each e ∈ E
SORT(E)                                   ▷ sorts edges by w* non-increasingly
for each e ∈ E do
    E* ← E* ∪ e
    if α(w*, E*) > A and β(w*, E*) < B then
        return True
    end if
end for
return False
```

Proof. 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. There are two cases: 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 1), Algorithm 1 will find E^*. In 2), 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 cannot do this anymore. At that point we will 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 α and β 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 α and β it is needed to iterate once again E (the part

in E^* for α, the part in $E \setminus E^*$ for β) 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 α and β 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

In Section 2.2 we developed Algorithm 1 to solve the problem of finding the Core Structural Connectivity Network in polynomial time. However, this algorithm does not guarantee a connected result. We solve the original problem, presented in Section 2, by first applying Algorithm 1 and then modifying the resulting core graph G^* to guarantee its connectedness. This results in an approximate solution for the full problem computable in polynomial time.

To extend G^* into a connected graph we add the necessary edges while decreasing the minimum possible the objective function f defined in Eq. 6. For this, we use the same approach that Wassermann et al. [10]. Namely, we make a multigraph G_{cc} with the connected components of G^* as nodes, complete it with all the possible edges between those connected components, and run a Maximum Spanning Tree algorithm. This selects the edges needed to produce a connected subgraph with the maximum possible weight. For the full details, see Wassermann et al. [10]. This way we get a connected subgraph close to the best possible subgraph, which we obtained using Algorithm 1.

3 Experiments and Results

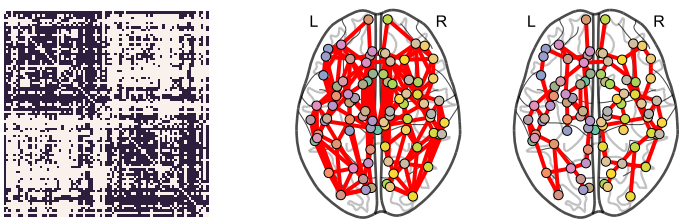

Fig. 1. Core structural connectivity network computed by our approach. On the left, we show the adjacency matrix for $\lambda = 0.5$, where 48.19% of the connections were included in the CSCN. In the central and right panels, we show the resulting CSCN for $\lambda = 0.9$ and 0.99 respectively. The percentage of included connections in the CSCN is 5.99% and 1.27% respectively.

We formalized the CSCN problem in section 2 and designed an algorithm to solve it in section 2.3. Now we will asses the performance of our method. For

this, we compare it with the most used [7] and with the recent one [10] in the task of connectivity prediction performance.

We use a subset of the HCP500 dataset [9]: all subjects aged 21-40 with complete dMRI protocol, totaling 309. We compute the weighted connectivity matrices between the cortical regions defined by the Desikan atlas [4] as done by Sotiropoulos et al. [9]. Examples of CSCN exctracted with our algorithm at different λ levels are shown in Fig. 1, which was generated using Nilearn [1].

3.1 Consistency of the Extracted Graph

To compare the stability across different algorithms for CSCN, we use an analysis based on Wassermann et al. [10]: we randomly take 500 subsets of 100 subjects each and computed the core graphs for all subsets. We then compute the number of *unstable connections*: connections that present in at least one core graph but not in all of them. Finally, in Table 1 we report each algorithm's *stability*:

$$\text{stability of the algorithm} \triangleq 1 - \frac{\#\{\text{unstable connections}\}}{\#\{\text{total connections}\}} \ .$$

This measure quantifies the CSCN consistency across subsamples. Due to the homogeneity of our sample, we expect the CSCNs obtained by an algorithm to be similar across subsamples. Hence, a stabler algorithm is preferable.

3.2 Predicting Handedness-Specific Connectivity

We evaluate performance of the methods by using the generated core graphs as a feature selection for handedness specific connectivity. We use 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.

Table 1. Stability of the algorithms and 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. [10], showing better statistical power. It's also more stable than Gong et al., showing improved consistency.

Algorithm	Features (mean)	Features (std)	Stability
Gong et al. 2009	0.066	0.256	0.364
Wassermann et al. 2016	0.415	0.723	0.644
Our approach ($\lambda = 0.50$)	1.042	1.269	0.528

Specifically, 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

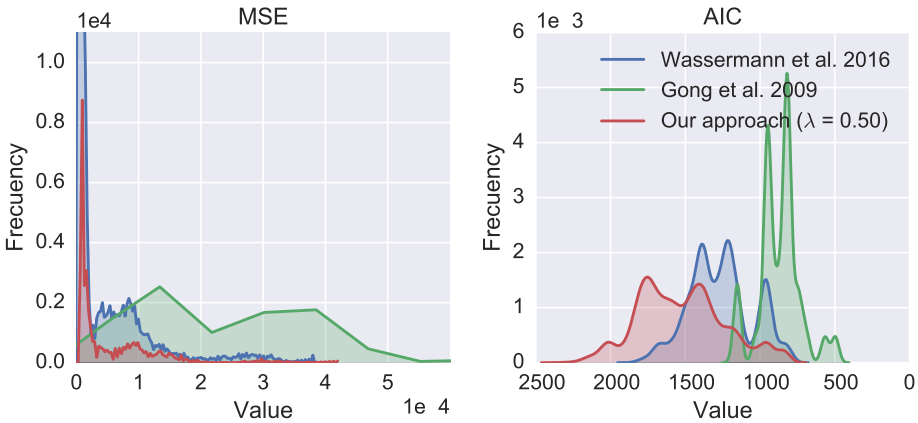

Fig. 2. Performance of core network as feature selection for a linear model for handedness specific connectivity. We evaluate model prediction (left) and fit (right) for Gong et al. [7] in green, Wassermann et al. [10] in blue and ours, in red. We show the histograms from our nested Leave-$\frac{1}{3}$-Out experiment. In both measures, our approach has more frequent lower values than Gong et al., showing a better performance.

correction for multiple hypotheses. This experiment is repeated 500 times. We quantify the amount of features that are selected after each procedure, which indicates how useful is the core graph algorithm for selecting the edges related to handedness. We show the results in Table 1.

To evaluate the prediction, we randomly take $\frac{1}{2}$ of the remaining subjects and fit a linear model on the features F to predict connectivity weights using the handedness of each subject. Finally, we predict the values of the features F from the handedness in the subjects left out. We quantify the quality of the linear model fitting Akaike Information Criterion (AIC) and of the prediction performance with the mean squared error (MSE) of the prediction. 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 results of this experiments in Fig. 2.

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.

Our results show that our algorithm outperforms, in the prediction experiment, the most used technique [7] as well as latest approaches [10]. In Table 1 we can see that our algorithm preserves, in average, more connections correlated with the handedness of the subjects. We can also see that despite being less

stable than Wassermann et al.'s it is stabler than Gong et al.'s. Finally, Fig. 2 shows that, in the handedness prediction experiment, our method outperforms Gong et al.'s and Wassermann et al's: the number of cases with lower AIC and MSE is larger in our case. Hence, our CSCN is better as linear model relating connectivity with handedness in terms of model fitting and prediction.

In terms of theoretical contributions, we formalized the problem, proved its 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. Moreover, our method avoids the double dipping problem by not choosing the feature set with hypothesis testing.

References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics 8(February), 14 (2014)
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. 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)
9. 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)
10. Wassermann, D., Mazauric, D., Gallardo Diez, G.A., Deriche, R.: Extracting the Core Structural Connectivity Network: Guaranteeing Network Connectedness Through a Graph-Theoretical Approach. In: MICCAI 2016 (2016)