

# Predicting Uncharted Connections with Szemerédi's Regularity Lemma in Cerebral Cortical Networks

T. Nepusz<sup>1,3</sup> F. Bacsó<sup>1</sup> L. Négyessy<sup>2</sup> G. Tusnády<sup>4</sup>

<sup>1</sup>Dept. of Biophysics  
KFKI Research Institute for Particle and Nuclear Physics

<sup>2</sup>Neurobiological Research Group  
Hungarian Academy of Sciences and Semmelweis University

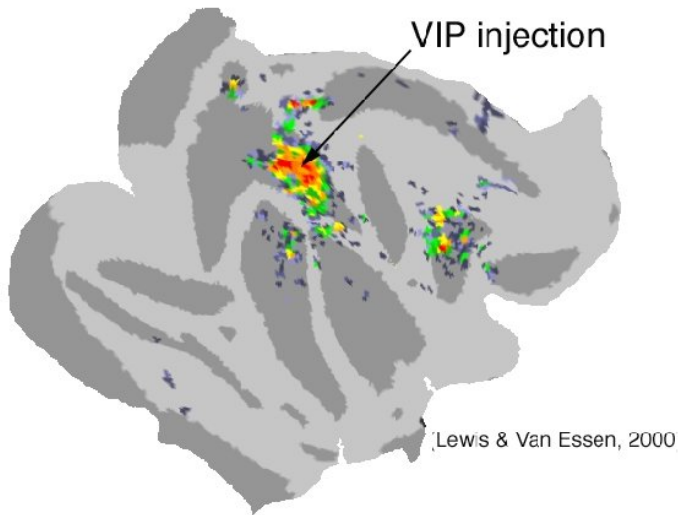
<sup>3</sup>Department of Measurement and Information Systems  
Budapest University of Technology and Economics

<sup>4</sup>Alfréd Rényi Institute of Mathematics  
Hungarian Academy of Sciences





# Structure of the Cerebral Cortex



# Graph Model of the Cerebral Cortex

- Areas of the cerebral cortex form a highly clustered, densely connected small-world network
- Nodes = different brain areas
- Edges = neuronal connections between areas
- **Problem:** not fully charted, some connections are still missing (methodological difficulties)
- Can we predict the unexplored connections from the known ones?

# Graph Model of the Cerebral Cortex

- Areas of the cerebral cortex form a highly clustered, densely connected small-world network
- Nodes = different brain areas
- Edges = neuronal connections between areas
- **Problem:** not fully charted, some connections are still missing (methodological difficulties)
- Can we predict the unexplored connections from the known ones?

# The Network That We Studied

- Macaque monkey
- Visual and sensorimotor cortex
- 45 areas, 463 **directed** connections (density = 0.23)
  - Visual: 30 areas, 335 connections (density = 0.39)
  - Sensorimotor: 15 areas, 85 connections (density = 0.4)
  - 43 connections are between visual and sensorimotor areas
- Missing edges can mean two different things:
  - 1 Either they are checked and found non-existent
  - 2 ...or they **might** be there – we just simply have no information regarding their existence!

# The Network That We Studied

- Macaque monkey
- Visual and sensorimotor cortex
- 45 areas, 463 **directed** connections (density = 0.23)
  - Visual: 30 areas, 335 connections (density = 0.39)
  - Sensorimotor: 15 areas, 85 connections (density = 0.4)
  - 43 connections are between visual and sensorimotor areas
- Missing edges can mean two different things:
  - 1 Either they are checked and found non-existent
  - 2 ...or they **might** be there – we just simply have no information regarding their existence!



# Szemerédi's Regularity Lemma

- **Roughly speaking**, Szemerédi's Regularity Lemma states that the node set of every **large** graph can be partitioned into a small number of disjoint subsets so that the subgraphs between any two of the subsets are “random-like”.
- Let us assume that the node set  $V$  of a graph  $G = (V, E)$  is partitioned into subsets  $V_1, V_2, \dots, V_n$ .
- According to the lemma, for any  $1 \leq i, j \leq n$ , the edges between  $V_i$  and  $V_j$  form a random-like subgraph with a connection probability  $p_{i,j}$ .

# Szemerédi's Regularity Lemma

- **Roughly speaking**, Szemerédi's Regularity Lemma states that the node set of every **large** graph can be partitioned into a small number of disjoint subsets so that the subgraphs between any two of the subsets are “random-like”.
- Let us assume that the node set  $V$  of a graph  $G = (V, E)$  is partitioned into subsets  $V_1, V_2, \dots, V_n$ .
- According to the lemma, for any  $1 \leq i, j \leq n$ , the edges between  $V_i$  and  $V_j$  form a random-like subgraph with a connection probability  $p_{i,j}$

# Szemerédi's Regularity Lemma – Directed Case

- In the case of directed graphs, we have **two** partitions instead of one:  $U_1, U_2, \dots, U_n$  and  $V_1, V_2, \dots, V_m$
- Roughly speaking,  $U$  partitions the nodes based on their *outgoing*, while  $V$  partitions them based on their *incoming* connections
- $p_{i,j}$  denotes the probability that an edge goes from a randomly chosen node from  $U_i$  to a randomly chosen node from  $V_j$
- E.g.  $p_{3,2} = 0.9$  means that 90% of the edges are present between nodes of  $U_3$  and  $V_2$  (directed towards  $V_2$ )

# Szemerédi's Regularity Lemma – Directed Case

- In the case of directed graphs, we have **two** partitions instead of one:  $U_1, U_2, \dots, U_n$  and  $V_1, V_2, \dots, V_m$
- Roughly speaking,  $U$  partitions the nodes based on their *outgoing*, while  $V$  partitions them based on their *incoming* connections
- $p_{i,j}$  denotes the probability that an edge goes from a randomly chosen node from  $U_i$  to a randomly chosen node from  $V_j$
- E.g.  $p_{3,2} = 0.9$  means that 90% of the edges are present between nodes of  $U_3$  and  $V_2$  (directed towards  $V_2$ )

# An Important Thing

- Szemerédi's Lemma...
  - ...is only an existence lemma (no algorithm)
  - ...only applies for **large** graphs (and our graph is definitely not large)
- How can it still be of some use to us?
- We use the partitions and the probability matrix as a **model** for our graph

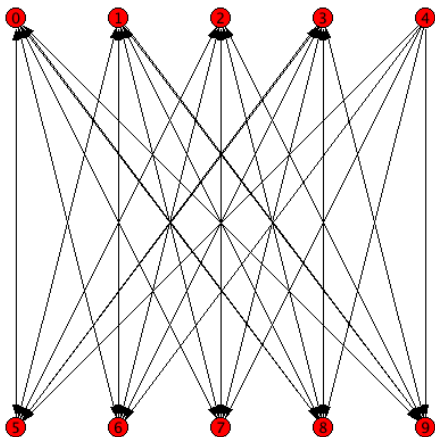
# An Important Thing

- Szemerédi's Lemma...
  - ...is only an existence lemma (no algorithm)
  - ...only applies for **large** graphs (and our graph is definitely not large)
- How can it still be of some use to us?
- We use the partitions and the probability matrix as a **model** for our graph

# An Important Thing

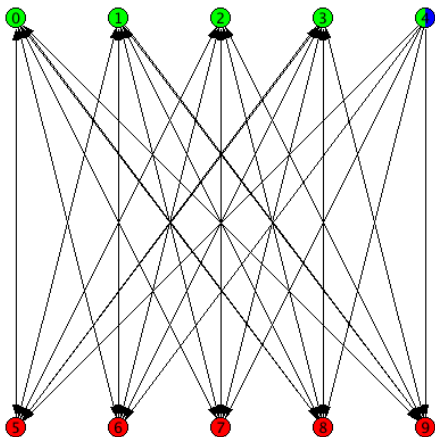
- Szemerédi's Lemma...
  - ...is only an existence lemma (no algorithm)
  - ...only applies for **large** graphs (and our graph is definitely not large)
- How can it still be of some use to us?
- We use the partitions and the probability matrix as a **model** for our graph

# A Simple Example



Note that node 4 has only outgoing edges!

## A Simple Example



$$U_1 = \{0, 1, 2, 3, 4\}$$

$$U_2 = \{5, 6, 7, 8, 9\}$$

$$V_1 = \{0, 1, 2, 3\}$$

$$V_2 = \{5, 6, 7, 8, 9\}$$

$$V_3 = \{4\}$$

$$\mathbf{P} = \begin{bmatrix} \text{green} & \text{red} \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{array}{l} \text{green} \\ \text{red} \\ \text{blue} \end{array}$$

Note that node 4 has only outgoing edges!



# Szemerédi's Lemma as a Model

- Can we use the lemma to predict possibly existing connections?
- Basic idea:
  - Let's assign the two "colors" of the nodes in some way (initial coloring)
  - Calculate the probability matrix
  - Calculate the likelihood of our network, given these probabilities
  - Do some slight mutations on the original coloring until we reach a local optimum in the likelihood (simulated annealing)
  - At the local optimum, elements of the probability matrix close to 1 denote almost regular structures where possibly more connections exist

# Generalization of the Model

- This basic model didn't work very well, so we generalized it a little bit:
  - Instead of just two “colors”, we used two 4 dimensional vectors (representing the incoming and outgoing “connectivity patterns” of the area)
  - A connection was present between any two nodes if the scalar product of the **outgoing** vector of the first one and the **incoming** vector of the second one was above a given threshold (so their “connectivity patterns” match)
  - The optimization process was the same (but the vectors were mutated instead of the colors)
- Can it be expressed by the model of Bollobás?  
(yesterday's talk)

# Generalization of the Model

- This basic model didn't work very well, so we generalized it a little bit:
  - Instead of just two “colors”, we used two 4 dimensional vectors (representing the incoming and outgoing “connectivity patterns” of the area)
  - A connection was present between any two nodes if the scalar product of the **outgoing** vector of the first one and the **incoming** vector of the second one was above a given threshold (so their “connectivity patterns” match)
  - The optimization process was the same (but the vectors were mutated instead of the colors)
- Can it be expressed by the model of Bollobás?  
(yesterday's talk)

# Generalization of the Model

- This basic model didn't work very well, so we generalized it a little bit:
  - Instead of just two “colors”, we used two 4 dimensional vectors (representing the incoming and outgoing “connectivity patterns” of the area)
  - A connection was present between any two nodes if the scalar product of the **outgoing** vector of the first one and the **incoming** vector of the second one was above a given threshold (so their “connectivity patterns” match)
  - The optimization process was the same (but the vectors were mutated instead of the colors)
- Can it be expressed by the model of Bollobás?  
(yesterday's talk)

# Generalized model - Results

- 394 connections were predicted correctly with 100% confidence
- 47 connections were predicted as “almost sure”
  - 7 out of this 47 were not present in the original network.
- 73 connections were predicted as “might be there”
  - 45 out of this 73 were not present in the original network.
- $45 + 7 = 52$  potential connection candidates were found.  
Which ones should we choose?

# Generalized model - Results

- 394 connections were predicted correctly with 100% confidence
- 47 connections were predicted as “almost sure”
  - 7 out of this 47 were not present in the original network.
- 73 connections were predicted as “might be there”
  - 45 out of this 73 were not present in the original network.
- $45 + 7 = 52$  potential connection candidates were found.  
Which ones should we choose?

## Generalized model - Results

- 394 connections were predicted correctly with 100% confidence
- 47 connections were predicted as “almost sure”
  - 7 out of this 47 were not present in the original network.
- 73 connections were predicted as “might be there”
  - 45 out of this 73 were not present in the original network.
- 45 + 7 = 52 potential connection candidates were found.  
Which ones should we choose?

## Generalized model - Results

- 394 connections were predicted correctly with 100% confidence
- 47 connections were predicted as “almost sure”
  - 7 out of this 47 were not present in the original network.
- 73 connections were predicted as “might be there”
  - 45 out of this 73 were not present in the original network.
- $45 + 7 = 52$  potential connection candidates were found.  
Which ones should we choose?

# Generalized model - Results

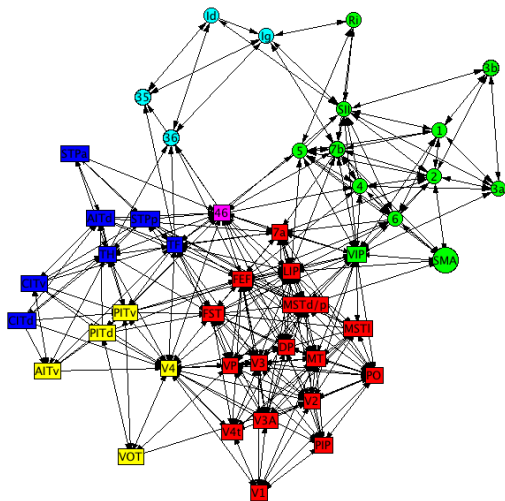
- Maintaining a connection in the cortex has a high metabolic cost, so we should choose the ones which
  - significantly decrease the sum of shortest path lengths while
  - they are predicted with a reasonably high level of confidence

# Generalized model - Results

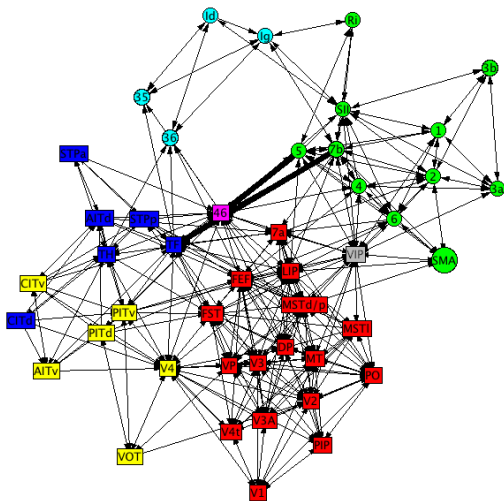
From	To	Confidence	$\Delta SP$	$\Delta SP \times \text{conf.}$
AITv	46	0.3836	-15	-5.7540
CITv	46	0.3836	-14	-5.3704
<b>5</b>	<b>46</b>	<b>0.8511</b>	<b>-4</b>	<b>-3.4044</b>
V4	STPa	0.3836	-7	-2.6852
V2	LIP	0.3836	-4	-1.5344
V4	AITd	0.3836	-4	-1.5344
V4	STPp	0.3836	-4	-1.5344
V4t	LIP	0.3836	-4	-1.5344
<b>7b</b>	<b>TF</b>	<b>0.3836</b>	<b>-3</b>	<b>-1.1508</b>

Highlighted connections are particularly important - multimodal connections have not been examined very thoroughly yet.

# Validity of Our Results



# Validity of Our Results



# Summary

- Szemerédi clustering not only represents the functionally relevant grouping of cortical areas, it is also a useful way to **identify areas of bridging position**.
- Although our method is able to point out yet uncharted connections worth testing by experimental methods, it also seems to be a **powerful technique in modeling the cortical network**. Accordingly, we were able to reproduce the cortical connectivity with high fidelity.
- Such models could be useful in understanding computation performed by the cortex.

# For Further Reading



E. Szemerédi

Regular Partitions of Graphs

*Problèmes combinatoires et théorie des graphes*, 399–401, 1978.



J. Komlós, M. Simonovits

Szemerédi's Regularity Lemma and its Applications in Graph Theory

*Combinatorics, Paul Erdős is Eighty*, 2:295–352, 1996



L. Négyessy, T. Nepusz, L. Kocsis, F. Bazsó

Prediction of the Main Cortical Areas and Connections Involved in the Tactile Function of the Visual Cortex by Network Analysis

*Eur. J. Neurosci.*, **23**(7):1919–1930, 2006.