Comparing meso-scale maps given by community detection algorithms on the Emmons laboratory *Caenorhabditis elegans* connectome data

George Vernon[a,1]

[a]Department of Physics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK

This manuscript was compiled on May 19, 2022

We apply three types of community detection to the the 2019 Emmons laboratory data (1) for the *Caenorhabditis elegans* (*C. elegans*) connectome and compare the results with those found by Pavlovic et al. in 2014 (2) for the same three types of community detection using the 2011 connectome data (3). Then we evaluate the three methods using prior biological knowledge about the *C. elegans* nervous system and discuss the implications of newly identified blocks arising from the fitting of the Erdős-Rényi Mixture Model (ERMM) (4) to the 2019 data.

C. Elegans | Connectome | Community detection | Neuroscience

Sydney Brenner proposed a scheme of research into *C. elegans* to the UK Medical Research Council in 1963, due to its simplicity, easiness to grow, and suitability for genetic analysis. It is now considered a model organism. He began recording the connections between its neurons with a microscope in the 1970s, and Olaf Sporns named this map of connections in matrix form a "connectome" in 2005 (5). The map consists of 302 neurons, and around 8,000 synapses. The complexity of the structure of this map and other complex systems necessitates a method for aggregating neurons into groups, for useful meso-scale visualisation and modelling. Community detection techniques provide one framework for doing this. There exist both deterministic and non-determistic algorithms for community detection: we consider the deterministic Fast Louvain and Spectral algorithms, and also the non-deterministic Erdős-Rényi Mixture Model (ERMM) (4), which is called a stochastic block model.

Stochastic block models are block models with the following property: consider a stochastic block model with K blocks and N nodes. Let g_i denote the block to which node i belongs, and then we can define a $K \times K$ matrix ψ such that $\psi_{g_i g_j}$ is the probability of a connection between nodes i and j. We are interested in the inverse problem: given the data, we want to find the number of blocks K and the matrix ψ that give the greatest likelihood of generating the data in hand.

White published the original *C. Elegans* connectome data from his electron micrography in 1986 (6). Varshney et al. (3) updated this data in 2011, using new electron micrography and unpublished laboratory data of White et al. to fill in some of the missing connections in the ventral cord. This still left the posterior ventral cord connection data incomplete. In 2019, Cook et al. (1) (re)annotated several thousand new and old images corresponding to the legacy hermaphrodite series N2U, N2T, JSE for the anterior and posterior, and N2T, N2W, and JSA for the pharynx. They used the software *Elegans* (7), and with it were able to find new synapses, give synaptic weights for the data, and fill in the missing connectivity for neurons in the posterior ventral cord.

A 2014 ERMM fit to the 2011 data in a block-modelling study of *C. Elegans* data by Pavlovic et al. (2) found nine distinct blocks, whose functions are broadly characterised through prior biological knowledge about their neurons: 1) chemosensation / thermosensation, 2) escape / avoidance, 3) motor posterior, 4) motor anterior, 5) command, 6) command, 7) unknown / egg-laying / defecation, 8) nose-touch / head motor, 9) motor anterior. Blocks 1, 2, and 9 are characterised by strong internal connections but weaker inter-modular connections. Pavlovic et al. note that posterior and anterior motor blocks 3 and 4 are distinguished by a strong lack of connectivity, but are at the time unsure whether this feature of the fit contains real biological data or is an artefact of the then-missing connection data for the posterior ventral region. They also question whether the ERMM categorization of neurons into block 9, which is combined with block 4 by both the Spectral and Louvain algorithms, is a real biological feature, or an artefact of missing data. We were motivated by the opportunities both to answer those questions and to obtain a more accurate meso-scale map of *C. Elegans* connectome to reapply the community detection methods originally chosen by Pavlovic et al. to the complete 2019 data.

Results. The ERMM, Louvain, and Spectral partitions are given in tables 1, 2, and 3 respectively. We discuss them each in turn. For brevity, we refer to L/R neurons belonging to separate groups as "orphaned", but we suggest that they may be correctly classified separately, exemplifying lateralisation of function. (8)

A. ERMM. The optimum ERMM for the normalized *C. Elegans* connectome data has an ICL of -9652, and eleven groups. Group 1 corresponds to the pharynx, including pharyngeal interneurons and motorneurons. Group 2 consists of 24 lateral

Significance Statement

We apply previously proven community detection algorithms to new data representing the *C. Elegans* neural wiring map. We present a new meso-scale map of the connectome with different neuronal groups than those offered before. The meso-scale map gives useful ways to understand input-response patterns at the circuits level, and may inform control-theoretic studies by clearly identifying the most isolated circuits in the worm.

[1]To whom correspondence should be addressed. E-mail: george.vernon@warwick.ac.uk

Table 1. The optimum ERMM fit parameters $\pi_{c_i c_j}$ gives the probability that neuron i connects to neuron j, with c_i the group indicator of neuron i, when adjacency A_{ij} is unknown. Values given correct to three decimal places.

	1	2	3	4	5	6	7	8	9	10	11
1	0.584	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000
2	0.000	0.377	0.000	0.207	0.022	0.000	0.000	0.059	0.011	0.049	0.022
3	0.000	0.000	0.224	0.000	0.011	0.055	0.034	0.020	0.000	0.000	0.291
4	0.000	0.207	0.000	0.360	0.104	0.001	0.007	0.180	0.057	0.130	0.327
5	0.000	0.022	0.011	0.104	0.210	0.008	0.019	0.095	0.015	0.151	0.258
6	0.000	0.000	0.055	0.001	0.008	0.289	0.001	0.253	0.000	0.000	0.290
7	0.000	0.000	0.034	0.007	0.019	0.001	0.236	0.089	0.000	0.001	0.330
8	0.000	0.059	0.020	0.180	0.095	0.253	0.089	0.582	0.015	0.035	0.458
9	0.005	0.011	0.000	0.057	0.015	0.000	0.000	0.015	0.257	0.156	0.079
10	0.000	0.049	0.000	0.130	0.151	0.000	0.001	0.035	0.156	0.393	0.156
11	0.000	0.022	0.291	0.327	0.258	0.290	0.330	0.458	0.079	0.156	0.833

ganglia interneurons, and two pairs of ventral ganglia interneurons, AIAx and AIYx. The neurons in group 2 are involved in sensation and the first layer of sensation integration. Group 3 consists of mostly posterior ventral cord motorneurons and proprioceptive / sensory motorneurons, and so it is implicated in motion. 20% of the group 4 neurons are orphans, and so it is less clearly important as a functional group. Its neurons belong to the lateral ganglia / head in location, and are sensory / integrative / regulatory in function. 28% of the neurons in group 5 are orphans, and so we interpret it as the weakest of the groups. It contains head interneurons, and connects strongly to group 10, actually possessing six partners of ventral ganglion orphans in group 10. Group 6 is strongly identified with no orphans. It contains tail/rear posterior and dorsorectal ganglion (DRG) neurons, whose functions are sensory / integratory / motor. Group 7 contains anterior motor / vulva eurons, involved in locomotion with some proprioceptive self-feedback through the VB and DB neurons (9). They also output to the VD neurons in this group. We label this group "anterior motor". Group 8 contains interneurons from a few ganglions, and some pioneering / guidepost neurons (AVG, PVPR, PVQR, PVT). The neurons in group 8 comprise the deepest brain region, and are implicated in defecation, egg-laying, and vulval function. We label it "deep brain". Group 9 contains neurons in the head and nerve ring. Their functions are sensory and sensory integratory. We label it "head nerve ring / labia". Group 10 contains head nerve ring interneurons, outer labia sensilia, and six orphans whose pairs are in group 5. (ADEL, second bulb; SIADR, SIAVL, SIBDR, SIBVR, SMBDR, ventral ganglion.) SMB neurons are known to set the amplitude of the sinusoid in locomotion (10). We label it "head nerve ring / outer labia / ventral ganglion". Finally, group 11 contains command interneurons involved in locomotion, that may be separately considered as drivers and modulators. Two neurons in group 11 are not drivers or modulators, these are DVA and PVR. DVA is involved in the touch circuit (11) and is itself a stretch receptor known to modulate locomotion (12). Its strong connectivity within group 11 and function as a stretch receptor suggests a role in involuntary (not intermediated by a ganglion) muscle reaction. The relevance of PVR is less clear and therefore very interesting. It is a stretch-sensitive interneuron belonging to the right lumbar ganglion, and has unclear function. We label group 11 "locomotion drive and modulation". The $\boldsymbol{\pi}$ table is given in Table 1.

B. Louvain. Group 1 is identical across the ERMM, Louvain, and Spectral fits.

Group 2 of the Louvain fit consists almost perfectly of group 2 and group 4 from the ERMM fit, giving almost the whole lateral ganglia. Only two pairs of symmetric neuron from group 2 or group 4 of the ERMM fit are not completely present in group 2 of the Louvain fit. The first is URXR, which is placed in group 3. URXR is implicated in a myriad of complex functions (13–16), which is likely related to its strong connections in both groups 2 and 3 of the Louvain fit, allowing it to integrate complex information. It makes 8 chemical synapses in block 2, 14 in block 3, and 4 in block 4. The second missing pair consists of ADAL and ADAR, which are both placed in group 3.

Group 3 of the Louvain fit contains URXR, ADAL, and ADAR, associated with group 2 as mentioned, and also contains group 10.

Group 4 of the Louvain fit contains half of group 11 (locomotion drive & modulation) from the ERMM fit, and most of group 7 (anterior motor + vulva) from the ERMM fit. It contains most of the drivers from ERMM group 7 but none of the modulators. It can be labelled "anterior motor and locomotion drive".

Group 5 of the Louvain fit contains the other half of group 11 from the ERMM fit (suggesting we might usefully conceive of group 11 of the ERMM as bridging groups 4 and 5 of Louvain), and most of group 3 from the ERMM fit. It contains the locomotion modulators AVDL, AVDR, DVA, PVCL and PVCR. It can be labelled "posterior motor and locomotion modulation".

C. Spectral. The Spectral fit is very similar to the Louvain fit, and it is appropriate to describe it in terms of its differences with the Louvain fit.

Discussion. Pavlovic et. al. write: "It is worth noting that the connectivity data for C. elegans are known to be partial or missing for 39 of 302 neurons, including 21 of the 75 locomotor motoneurons [63] and the data for the posterior parts of the nerve cords are especially sparse and uncertain. It is therefore unclear whether this split between Blocks 3, 4 and 9 contains biological information or whether a more complete mapping of connections in the posterior part of the ventral cord would alter these results."

Block 3 of our ERMM fit corresponds to Pavlovic et al.'s block 3 posterior motor, labelled by us as "posterior motor &

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Group 11
I1L	**ADFR**	AS06	ADAL	**ADER**	AS10	AS01	VB05	AVFL	**ALNL**	SMDVL	AVAL
I1R	AFDL	AS07	ADAR	ALA	AS11	AS02	VC01	AVFR	ILIDL	CEPDL	SMDVR	AVAR
I2L	AFDR	AS08	**ADFL**	ALML	DA08	AS03	VC02	AVG	ILIDR	CEPDR	**URBR**	AVBL
I2R	AIAL	AS09	ADLL	ALMR	DA09	AS04	VC03	AVHL	ILIL	CEPVL	URYVL	AVBR
I3	AIAR	DA07	ADLR	**ALNR**	DB07	AS05	VD01	AVHR	ILIR	CEPVR	URYVR	AVDL
I4	AINL	DB04	AIBL	AVKL	DD06	DA01	VD02	AVJL	ILIVL	**OLLR**		AVDR
I5	AINR	DB05	AIBR	AVKR	DVB	DA02	VD03	AVJR	ILIVR	RIAL		AVEL
I6	AIYL	DB06	AIML	AVM	LUAL	DA03	VD04	AVL	IL2DL	RIAR		AVER
M1	AIYR	DD04	AIMR	BDUL	LUAR	DA04	VD05	DVC	IL2DR	RIBL		DVA
M2L	ASEL	DD05	AIZL	BDUR	PDA	DA05	VD06	**HSNR**	IL2L	RIBR		PVCL
M2R	ASER	PDEL	AIZR	**FLPL**	PDB	DA06		PVNL	IL2R	RICL		PLCR
M3L	ASGL	PDER	AQR	PLNL	PHAL	DB01		PVNR	IL2VL	RICR		PVR
M3R	ASGR	PLML	ASHL	PLNR	PHAR	DB02		PVPL	IL2VR	RIH		
M4	ASIL	PLMR	ASHR	**RIFL**	PHBL	DB03		PVPR	**OLLL**	RIS		
M5	ASIR	**PVDR**	**AUAL**	RIML	PHBR	DD01		PVQL	OLQDL	RIVL		
MCL	ASJL	PVM	BAGL	RIMR	PHCL	DD02		PVQR	OLQDR	RIVR		
MCR	ASJR	VA08	BAGR	RMFL	PHCR	DD03		RVT	OLQVL	RMDDL		
MI	ASKL	VA09	**HSNL**	RMFR	PQR	**FLPR**		RID	OLQVR	RMDDR		
NMSL	ASKR	VA10	**RIFR**	SAADL	PVWL	**PVDL**			RIPL	RMDL		
NSMR	**AUAR**	VA11	RIGL	SAADR	PVWR	SABD			RIPR	RMDR		
	AWAL	VB06	RIGR	SAAVL	VA12	SABVL			RMED	RMDVL		
	AWAR	VB07	RIR	SAAVR	VB11	SABVR			RMEL	RMDVR		
	AWBL	VB08	**RMGR**	SDQL	VC04	VA01			RMER	**RMGL**		
	AWBR	VB09	URXL	SDQR	VC05	VA02			RMEV	RMHL		
	AWCL	VB10	URXR	**SIADL**	VC06	VA03			URADL	RMHR		
	AWCR	VD07		**SIAVR**	VD11	VA04			URADR	**SIADR**		
		VD08		**SIBDL**	VD12	VA05			URAVL	**SIAVL**		
		VD09		**SIBVL**	VD13	VA06			URAVR	**SIBDR**		
		VD10		**SMBDL**		VA07			**URBL**	**SIBVR**		
				SMBVL		VB02			URYDL	**SMBDR**		
				SMBVR		VB03			URYDR	SMDDL		
				VB01		VB04				SMDDR		

Fig. 1. The optimum ERMM groups. Bold neurons identify symmetric neurons whose left/right (L/R) pairs are split into two groups.

Group 1	Group 2		Group 3		Group 4			Group 5		
I1L	ADFL	AUAR	ADAL	PLNR	SAAVR	AS01	VA03	**ALML**	PDA	VA12
I1R	ADFR	AVFL	ADAR	**RIAR**	**SDQR**	AS02	VA04	AQR	PDB	VB06
I2L	ADLL	AVFR	ADEL	**RIBR**	SIADL	AS03	VA05	AS07	PDEL	VB07
I2R	ADLR	AWAL	ADER	RICL	SIADR	AS04	VA06	AS08	PDER	VB08
I3	AFDL	AWAR	**ALMR**	RICR	SIAVL	AS05	VA07	AS09	PHAL	VB09
I4	AFDR	AWBL	ALNL	RIH	SIAVR	AS06	VB01	AS10	PHAR	VB10
I5	AIAL	AWBR	ALNR	RIML	SIBDL	**AVAR**	VB02	AS11	PHBL	VB11
I6	AIAR	AWCL	AVKL	RIMR	SIBDR	AVBL	VB03	**AVAL**	PHBR	VC06
M1	AIBL	AWCR	AVKR	RIPL	SIBVL	AVBR	VB04	AVDL	PHCL	VD07
M2L	AIBR	BAGL	CEPDL	RIPR	SIBVR	AVEL	VB05	AVDR	PHCR	VD08
M2R	AIML	BAGR	CEPDR	RIS	SMBDL	AVER	VC01	AVG	PLML	VD09
M3L	AIMR	HSNL	CEPVL	RIVL	SMBDR	DA01	VC02	AVHL	PLMR	VD10
M3R	AINL	HSNR	CEPVR	RIVR	SMBVL	DA02	VC03	AVHR	PQR	VD11
M4	AINR	PVQL	IL1DL	RMDDL	SMBVR	DA03	VD01	AVJL	PVCL	VD12
M5	AIYL	PVQR	IL1DR	RMDDR	SMDDL	DA04	VD02	AVJR	PVCR	VD13
MCL	AIYR	**RIAL**	IL1L	RMDL	SMDDR	DA05	VD03	AVL	PVDL	
MCR	AIZL	**RIBL**	IL1R	RMDR	SMDVL	DA06	VD04	AVM	PVDR	
MI	AIZR	RIFL	IL1VL	RMDVL	SMDVR	DB01	VD05	BDUL	PVM	
NSML	ALA	RIFR	IL1VR	RMDVR	URADL	DB02	VD06	BDUR	PVNL	
NSMR	ASEL	RIGL	IL2DL	RMED	URADR	DB03		DA07	PVNR	
	ASER	RIGR	IL2DR	RMEL	URAVL	DB04		DA08	PVPL	
	ASGL	RIR	IL2L	RMER	URAVR	DB05		DA09	PVPR	
	ASGR	**URXL**	IL2R	RMEV	URBL	DD01		DB06	PVR	
	ASHL	VC04	IL2VL	RMFL	URBR	DD02		DB07	PVT	
	ASHR	VC05	IL2VR	RMFR	**URXR**	DD03		DD04	PVWL	
	ASIL		OLLL	RMGL	URYDL	FLPL		DD05	PVWR	
	ASIR		OLLR	RMGR	URYDR	FLPR		DD06	SABD	
	ASJL		OLQDL	RMHL	URYVL	RID		DVA	**SDQL**	
	ASJR		OLQDR	RMHR	URYVR	SABVL		DVB	VA08	
	ASKL		OLQVL	SAADL		SABVR		DVC	VA09	
	ASKR		OLQVR	SAADR		VA01		LUAL	VA10	
	AUAL		PLNL	SAAVL		VA02		LUAR	VA11	

Fig. 2. The optimum Louvain groups. Bold neurons identify symmetric neurons whose left/right (L/R) pairs are split into two groups.

Group 1	Group 2		Group 3			Group 4			Group 5	
I1L	ADAL	AWBL	ADEL	PLNL	RMHL	URYVL	AIML	PHBR	AQR	DD01	VA10
I1R	ADAR	AWBR	ADER	PLNR	RMHR	URYVR	AIMR	PHCL	AS01	DD02	VA11
I2L	ADFL	AWCL	**ALMR**	RIAL	SAADL		ALA	PHCR	AS02	DD03	VB01
I2R	ADFR	AWCR	ALNL	RIAR	SAADR		**ALML**	PQR	AS03	DD04	VB02
I3	ADLL	RIR	ALNR	RIBL	SAAVL		AS11	PVNL	AS04	DD05	VB03
I4	ADLR	**URXL**	**AVER**	RIBR	SAAVR		ASJL	PVNR	AS05	DVA	VB04
I5	AFDL		AVKL	RICL	**SDQR**		ASJR	PVPL	AS06	FLPL	VB05
I6	AFDR		AVKR	RICR	SIADL		AVFL	PVPR	AS07	FLPR	VB06
M1	AIAL		BAGL	RIGL	SIADR		AVFR	PVQL	AS08	PDEL	VB07
M2L	AIAR		BAGR	RIGR	SIAVL		AVG	PVQR	AS09	PDER	VB08
M2R	AIBL		CEPDL	RIH	SIAVR		AVHL	PVT	AS10	PLML	VB09
M3L	AIBR		CEPDR	RIML	SIBDL		AVHR	PVWL	AVAL	PLMR	VB10
M3R	AINL		CEPVL	RIMR	SIBDR		AVJL	PVWR	AVAR	PVCL	VC01
M4	AINR		CEPVR	RIPL	SIBVL		AVJR	RIFL	AVBL	PVCR	VC02
M5	AIYL		IL1DL	RIPR	SIBVR		AVL	RIFR	AVBR	PVDL	VC03
MCL	AIYR		IL1DR	RIS	SMBDL		BDUL	VA12	AVDL	PVDR	VD01
MCR	AIZL		IL1L	RIVL	SMBDR		BDUR	VB11	AVDR	PVM	VD02
MI	AIZR		IL1R	RIVR	SMBVL		DA08	VC04	**AVEL**	PVR	VD03
NSML	ASEL		IL1VL	RMDDL	SMBVR		DA09	VC05	AVM	RID	VD04
NSMR	ASER		IL1VR	RMDDR	SMDDL		DB07	VC06	DA01	SABD	VD05
	ASGL		IL2DL	RMDL	SMDDR		DD06	VD11	DA02	SABVL	VD06
	ASGR		IL2DR	RMDR	SMDVL		DVB	VD12	DA03	SABVR	VD07
	ASHL		IL2L	RMDVL	SMDVR		DVC	VD13	DA04	**SDQL**	VD08
	ASHR		IL2R	RMDVR	URADL		HSNL		DA05	VA01	VD09
	ASIL		IL2VL	RMED	URADR		HSNR		DA06	VA02	VD10
	ASIR		IL2VR	RMEL	URAVL		LUAL		DA07	VA03	
	ASKL		OLLL	RMER	URAVR		LUAR		DB01	VA04	
	ASKR		OLLR	RMEV	URBL		PDA		DB02	VA05	
	AUAL		OLQDL	RMFL	URBR		PDB		DB03	VA06	
	AUAR		OLQDR	RMFR	**URXR**		PHAL		DB04	VA07	
	AWAL		OLQVL	RMGL	URYDL		PHAR		DB05	VA08	
	AWAR		OLQVR	RMGR	URYDR		PHBL		DB06	VA09	

Fig. 3. The optimum Spectral groups. Bold neurons identify symmetric neurons whose left/right (L/R) pairs are split into two groups.

Fig. 4. Visualisation of the rearranged adjacency matrix with each of the three community detection methods: A: Louvain, B: Spectral, C: ERMM.

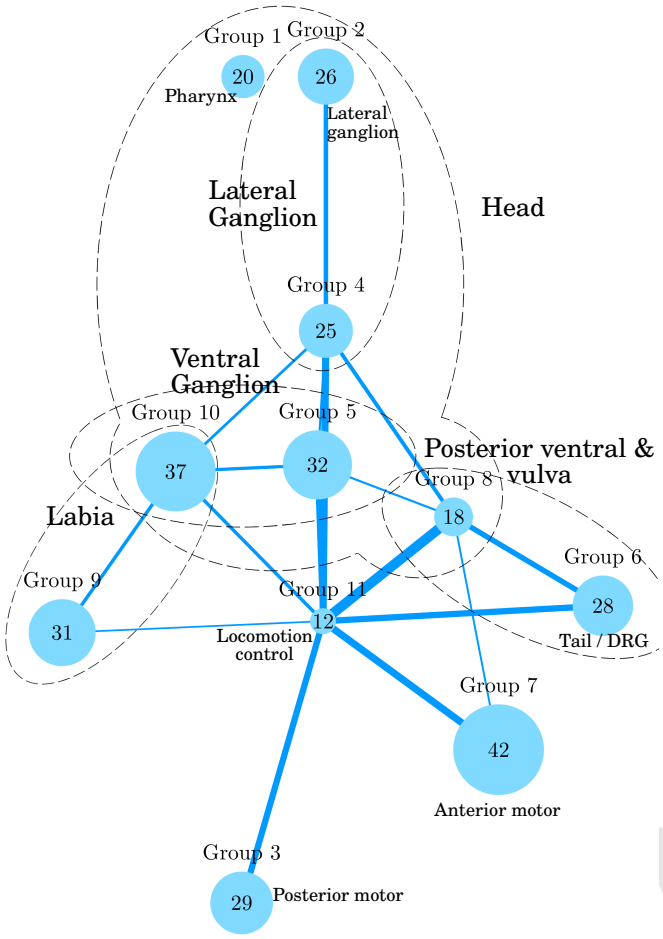

Fig. 5. Labeled visualisation of the group connection probabilities in the ERMM fit. Arc widths are proportional to connection probability, and the nodes representing each group are sized proportionally to the number of neurons they contain. Connection probabilities < 7% have been removed. The number written inside each node gives the number of neurons contained in that group.

proprioception", and block 7 of our ERMM fit corresponds to Pavlovic et al.'s blocks 4 & 9, labelled by them as anterior motor and labelled by us as "anterior motor & vulva". This distinction persists after repeating their method with complete data, and so we conclude that the split between the anterior and posterior motor blocks in *C. Elegans* does contain biological information. In other words, we are justified to consider the anterior and posterior motor groups as separate on the mesoscale. Pavlovic et al.'s block 9 consists of RVG and ventral cord neurons implicated in locomotion. The Louvain algorithm groups these together entirely as a subset of group 5, while the Spectral algorithm splits this group up entirely. The ERMM, on the other hand, groups the neurons from Pavlovic et al.'s ERMM blocks 4 and 9 together into group 7. A reasonable interpretation is that the complete connectivity data shows the original disconnectedness of those two blocks to be due to missing data, rather than representing a biological distinction.

The 2019 Emmons laboratory data gives a much more complete mapping of connections in the posterior ventral cord, and the optimal ERMM fit overall gives a very different set of mappings than that found on the incomplete data by Pavlovic et al. Block 2 is distinguished from block 4 mostly due block 4's connections with block 11. (Block 2 and block 4 have 2% and 33% connection probabilities with block 11 respectively.) Overall, the ERMM fit separates more L & R symmetric neurons: the Spectral algorithm separates four pairs, the Louvain algorithm separates six pairs, and the ERMM separates sixteen pairs. These were interpreted in (2) as misclassifications, however we could also view this as a hypothesis about lateralisation of function. Group 5 of the ERMM fit has the greatest number of separated symmetric neurons, and the greatest proportion of separated symmetric neurons — at 28% — with nine of its neurons having their symmetric pairs in other groups. Should group 5 of neurons be found to have a shared assymetric function, then the separation of the left and right pairs by the ERMM would be justified. A correct prediction about function would be remarkable given the data are structural only, and we are not considering directionality.

In this paper we haven't evaluated the goodness of the ERMM fit against prior biological information using a Rand measure as done in (2), and doing so would be a natural extension of this work.

It would be interesting to explore the application of a tiered stochastic block model to the *C. Elegans* data, applying the ERMM recursively to the blocks identified in the previous iteration, and then optimising over both parent and child ICLs. For example, blocks 9 and 10 in the ERMM sparsity matrix shown in Fig. 4 have very similar connectivity patterns, and look as though they could together comprise a larger "parent" block.

Finally, we note that while a directionless, binary graph is useful to generate the meso-scale map, once we have the map it may be informative to reconsider the fine-grained connection data and attempt to label the meso-scale map with inputs and outputs to better understand the worm. Future research should also implement more of the available data for the vertices and edges (developmental age, cell type, function, location, synapse type, etc.) in producing useful meso-scale maps.

Materials and Methods

We used the 2019 Emmons laboratory data (1) for the hermaphrodite, combining chemical synapse and gap junction adjacencies and manipulating them into a symmetric, binary adjacency matrix A of size 300×300 and 7064 non-zero values such that $A_{ij} = 1$ if there exists at least one chemical synapse or gap junction between neurons i and j, and $A_{ij} = 0$ otherwise. The two neurons not present in the adjacency matrix are CANL and CANR, which do not have chemical synapses nor gap junctions with other neurons. Their function is unknown but they are essential for survival of the worm. (10)

We now describe the three community-detection algorithms which we used: the ERMM, Spectral algorithm, and Fast Louvain algorithm.

ERMM. In contrast to traditional community detection algorithms, the ERMM (4) maximizes not only diagonal in-block connections but also maximizes off-diagonal block connections, discriminating between groups of highly connected neurons by their different connectivity with other groups.

Vertices belong each to one of Q classes, with prior probabilities $\alpha_1, ..., \alpha_Q$. We use the indicator variables Z_{iq} with $\sum Z_{iq} = 1$. Then $\alpha_q = P\{Z_{iq} = 1\} = P\{i \in q\}$ with $\sum_q \alpha_q = 1$. Denote by π_{ql} the probability that a vertex from class q connects with a vertex from class l. Because our graph is unconnected, we have $\pi_{ql} = \pi_{lq}$.

Then, suppose edges $\{X_{ij}\}$ are conditionally independent given the classes of i and j:

$$\begin{cases} X_{ij}|\{i \in q, j \in l\} \sim \mathcal{B}(\pi_{ql}) & \text{for } i \neq j \\ X_{ii} = 0 \end{cases}$$

Note $\{Z_{iq}\} \iff \{i \in q\}$. We have now fully described the ERMM. The MixeR package for R (4, 17–20) attempts to fit the optimal ERMM parameters to given data, using the EM algorithm (21). As is usual with incomplete data problems, the likelihood is intractable, but Daudin et al. (4) give a lower bound on $log\mathcal{L}(\mathcal{X})$ which is optimisable:

$$\mathcal{J}(R_{\mathcal{X}}) = \log \mathcal{L}(\mathcal{X}) - \mathrm{KL}[R_{\mathcal{X}}(\cdot), \mathrm{P}(\cdot|\mathcal{X})], \quad [1]$$

where KL is the Kullback-Leibler divergence and $\mathrm{P}(\mathcal{Z}|\mathcal{X})$ is the true conditional distribution of the indicators $\mathcal{Z}$ depending on the data $\mathcal{X}$, and $\mathcal{X}$ is an approximation of this conditional distribution, depending on $\mathcal{X}$.

Daudin et al. then give the estimation algorithm:

$$\hat{\alpha_q} = \frac{1}{n}\sum_{i=1}^{n} \tau_{iq}, \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \hat{\tau}_{iq}\hat{\tau}_{jl}x_{ij}}{\sum_{i \neq j} \hat{\tau}_{iq}\hat{\tau}_{jl}} \quad [2]$$

$$\hat{\tau}_{iq} \propto \prod_{i \neq j}\prod_{l} [\hat{\pi}_{ql}^{x_{ij}}(1-\hat{\pi}_{ql})^{1-x_{ij}}]^{\hat{\tau}_{jl}} \quad [3]$$

Where x_ij is the observation from the random data X_{ij}. Then, the classification of a node is given by

$$\hat{Z}_{iq} = \begin{cases} 1 & q = \mathrm{argmax}_{q'}(\hat{\tau}_{iq'}) \\ 0 & \text{otherwise} \end{cases} \quad [4]$$

where $\hat{\boldsymbol{\tau}}_i = (\hat{\tau}_{1i}, ..., \hat{\tau}_{1Q})$.

This optimization of the lower bound of the log likelihood is used by MixeR to calculate the optimum ERMM parameters given the number of classes Q. To compare across Q, we use the *Integrated Classification Likelihood* (ICL).

Integrated Classification Likelihood. For a model $\mathcal{M}$ with Q classes, the ICL is given by:

$\mathrm{ICL}(\mathcal{M}_Q) = \max \log \mathcal{L}(\mathcal{X}, \hat{\mathcal{Z}}|\psi, \mathcal{M}_Q)$
$- \frac{1}{2} \times \frac{Q(Q+1)}{2} \log(\frac{n(n-1)}{2}) - \frac{Q=1}{2}\log(n)$,

where $\hat{\mathcal{Z}}$ is the prediction of the unknown $\mathcal{Z}$ that generated the data $\mathcal{X}$, and $\psi = \{\boldsymbol{\alpha}, \boldsymbol{\pi}\}$ are the model parameters, and $\log \mathcal{L}(\mathcal{X}, \hat{\mathcal{Z}}|\psi, \mathcal{M}_Q)$ is the complete data log likelihood. (4)

The first term measures the clustering, and the negative terms penalise the complexity of the explanatory model.

Modularity. Both the Spectral and Fast Louvain algorithms are deterministic modularity-maximizing algorithms, and so before writing those algorithms we will first define the modularity of a graph partition.

The modularity is an objective function which assesses the goodness of a graph partition, and is written:

$$f_M = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{\rho(v_i)\rho(v_j)}{2m}\right)\delta(c_i, c_j) \quad [5]$$

where m is the number of edges, A_{ij} is the adjacency of vertices i and j in the observed data, and $\rho(v_i)$ is the expectation that vertex i makes a connection with any other given vertex according to the null model, c_i is the group indicator of vertex i st. $c_i = k$ if $v_i \in k$, and δ is the Kronecker-delta.

The null model used for the fast Louvain and Spectral algorithms is that edges are distributed at random in an equivalent graph with m edges. This gives $\rho_i\rho_j = k_ik_j$ where k_i is the degree of vertex i observed in the data.

The modularity can be understood as rewarding partitions which have higher in-group connectedness than expected (driven by the first term in the sum), and penalising partitions which have lower in-group connectedness than expected (driven by the second term in the sum) (22–24). The modularity does not explicitly measure between-group connectedness, but maximizing the modularity acts so as to minimize connectedness between groups.

Spectral. The Spectral algorithm could be considered to originate with Fiedler (25). It involves beginning with a modularity-optimizing bipartition of the graph, and then making more modularity-optimizing bipartitions of each partition until no further bipartition will give an improvement to the modularity. It can be carried out in a few equivalent ways; we will describe the eigenvector method for a graph with m edges and N vertices (26).

We use indicator variables $\boldsymbol{s}$ with s_i given by:

$$s_i = \begin{cases} +1 & \text{group 1} \\ -1 & \text{group 2} \end{cases} \quad [6]$$

satisfying $\delta(c_i, c_j) = (s_is_j + 1)/2$.

Then, the modularity of partition $\boldsymbol{s}$ is given:

$$f_M = \frac{1}{2}\sum_{i,j}[A_{ij} - P_{ij}](s_is_j + 1) = \frac{1}{2}\sum_{i,j}[A_{ij} - P_{ij}](s_is_j) \quad [7]$$

i.e.,

$$f_M = \frac{\boldsymbol{s}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{s}}{2}, \quad [8]$$

where $\boldsymbol{A}-\boldsymbol{P} = \boldsymbol{D}$. Modularity matrix $\boldsymbol{D}$ is symmetric, and therefore has N orthonormal eigenvectors $\boldsymbol{u}_i$, so we can write:

$$\boldsymbol{s} = \sum_{i}^{N} a_i\boldsymbol{u}_i, \quad [9]$$

where $a_i = \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{s}$ and we denote with β_i the eigenvalue corresponding to normalized eigenvector $\boldsymbol{u}_i$. Then we have:

$$f_M = \sum_i \frac{a_i^2\beta_i}{2}. \quad [10]$$

Relabelling the eigenvalues s.t. $\beta_1 \leq \beta_2 \leq ..... \leq \beta_N$, modularity maximization is seen to involve placing as much weight as possible on the first a_i in the sum. The heuristic is that the s_i are chosen so as to maximize $\boldsymbol{a_1} = \boldsymbol{u_1}^{\mathrm{T}}\boldsymbol{s}$:

$$s_i = \begin{cases} +1 & u_{1i} \geq 0 \\ -1 & u_{1i} < 0 \end{cases} \quad [11]$$

The algorithm is repeatedly iterated over each group until no new partitions can be made to increase the modularity.

Fast Louvain. The fast Louvain algorithm was proposed by Blondel et al. in 2008 (27). The heuristic operates in two stages to maximize the modularity. Initially, there is a random initialization of labels, and each vertex in the network is assigned its own group. Then, in the first stage, for each vertex V_i, each of its neighbours V_j are considered, and the modularity gain of placing vertex i in c_j is calculated. Vertex V_i is then placed in the group with the greatest modularity gain, or left in its own group is no positive modularity gain is possible. In the second stage of the algorithm, a new set of vertices is initialized such that each vertex represents one of the groups found. These two stages are then repeated until no more modularity gain is possible.

Computation. The optimum Louvain and Spectral block models were computed by taking the fits with the highest modularity from 20,000 runs using the Brain Connectivity Toolbox (28) in Matlab. As reported in (2), we found that 20,000 runs reliably gave the optimum fit each time. These 20,000 iterations took $\sim$ 15 minutes of computational time each, using an i5-7200U @ 2.50 GHz with 16 GB. The reason for taking the maximum of 20,000 iterations of a deterministic algorithm is that numerical discrepancies may occur in the permutations stage of the Spectral algorithm, erroneously changing the sign of the u_i, though this was not observed. On the other hand, the fit given by the Fast Louvain algorithm depends on a random initialization at the beginning, and so 20,000 iterations are necessary to reliably find the best fit.

Also corresponding with (2), we used the MixeR package for R to compute the ERMM. First we computed 100 runs with `qmin = 2` and `qmax = 50` to confidently identify the peak with $Q \in \{5, ..., 15\}$. We used the mixer function with `qmin = 5`, `qmax = 15`, `nbiter = 80`, and `fpnbiter = 40`, and took the fit with the highest ICL

from 7,000 iterations, which is seven times what (2) reported was sufficient to revisit the optimum fit "multiple times". We revisited the optimum solution four times across 7,000 random restarts, which took $\sim$ 200 hours of computational time using an i5-7200U @ 2.50 GHz with 16 GB.

The Emmons laboratory connectome data used is archived at `wormwiring.org`, accessed 09/04/22.

Readers are not able to access our data in this paper because this is not a real paper, but were this a real paper, this statement would tell you where to access the complete data.

ACKNOWLEDGMENTS. The author thanks Vincent Miele for the C++ mixnet package with which to implement the ERMM, and the nonextant Statistics for Systems Biology group at the former Institut national de la recherche agronomique for providing the R wrapper called MixeR.

1. Cook SJ, et al. (2019) Whole-animal connectomes of both caenorhabditis elegans sexes. *Nature* 571(7763):63–71.
2. Pavlovic DM, Vértes PE, Bullmore ET, Schafer WR, Nichols TE (2014) Stochastic blockmodeling of the modules and core of the caenorhabditis elegans connectome. *PLoS ONE* 9(7).
3. Varshney LR, Chen BL, Paniagua E, Hall DH, Chklovskii DB (2011) Structural properties of the caenorhabditis elegans neuronal network. *PLoS Computational Biology* 7(2).
4. Daudin JJ, Picard F, Robin S (2008) A mixture model for random graphs. *Statistics and Computing* 18(2):173–183.
5. Sporns O, Tononi G, Kötter R (2005) The human connectome: A structural description of the human brain. *PLoS Computational Biology* 1(4):0245–0251.
6. White (1986) The structure of the nervous system of the nematode caenorhabditis elegans. *Phil. Trans. R. Soc. Lond. B* 314(1165):1–340.
7. Xu M, et al. (2013) Computer assisted assembly of connectomes from electron micrographs: Application to caenorhabditis elegans. *PLoS ONE* 8(1).
8. Hobert O, Johnston RJ, Chang S (2002) Left–right asymmetry in the nervous system: the caenorhabditis elegans model. *Nature Reviews Neuroscience* 3(8):629–640.
9. Wen Q, et al. (2012) Proprioceptive coupling within motor neurons drives c. elegans forward locomotion. *Neuron* 76(4):750–761.
10. Hall DH, Altun ZF (2008) *C. Elegans Atlas*. (Cold Spring Harbor Press), 1 edition.
11. Wicks SR, Roehrig CJ, Rankin CH (1996) A dynamic network simulation of the nematode tap withdrawal circuit: predictions concerning synaptic function using behavioral criteria. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 16(12):4017–4031.
12. Li W, Feng Z, Sternberg PW, Xu XZS (2006) A c. elegans stretch receptor neuron revealed by a mechanosensitive trp channel homologue. *Nature* 440(7084):684–687.
13. Chang AJ, Chronis N, Karow DS, Marletta MA, Bargmann CI (2006) A distributed chemosensory circuit for oxygen preference in c. elegans. *PLOS Biology* 4(9):1–15.
14. Chang AJ, Bargmann CI (2008) Hypoxia and the hif-1 transcriptional pathway reorganize a neuronal circuit for oxygen-dependent behavior in caenorhabditis elegans. *Proceedings of the National Academy of Sciences* 105(20):7321–7326.
15. Rogers C, Persson A, Cheung B, de Bono M (2006) Behavioral motifs and neural pathways coordinating responses and aggregation in c. elegans. *Current Biology* 16(7):649–659.
16. Zimmer M, et al. (2009) Neurons detect increases and decreases in oxygen levels using distinct guanylate cyclases. *Neuron* 61(6):865–879.
17. Zanghi H, Ambroise C, Miele V (2008) Fast online graph clustering via erdős–rényi mixture. *Pattern Recognition* 41(12):3592–3599.
18. Zanghi H, Picard F, Miele V, Ambroise C (2008) Strategies for online inference of network mixture, Technical report.
19. Zanghi H, Picard F, Miele V, Ambroise C (2010) Strategies for online inference of model-based clustering in large and growing networks. *The Annals of Applied Statistics* 4(2).
20. Latouche P, Birmelé E, Ambroise C (2012) Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling* 12(1):93–115.
21. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.
22. Luczak T (1989) Sparse random graphs with a given degree sequence in *Proceedings of the Symposium on Random Graphs, Poznan*. pp. 165–182.
23. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms* 6(2-3):161–180.
24. (2008) *Handbook of Probability: Theory and Applications AU - Pattison, Philippa AU - Robins, Garry*. (SAGE Publications, Inc., Thousand Oaks), pp. 291–312. 18 Probabilistic Network Analysis.
25. Fiedler M (1973) Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* 23:298–305.
26. Newman MEJ (2006) Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* 74(3).
27. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.
28. Rubinov M, Kötter R, Hagmann P, Sporns O (2009) Brain connectivity toolbox: a collection of complex network measurements and brain connectivity datasets. *Neuroimage* 47.