Growing Hierarchical Self Organizing Maps for Community Detection

David McDonald Shan He

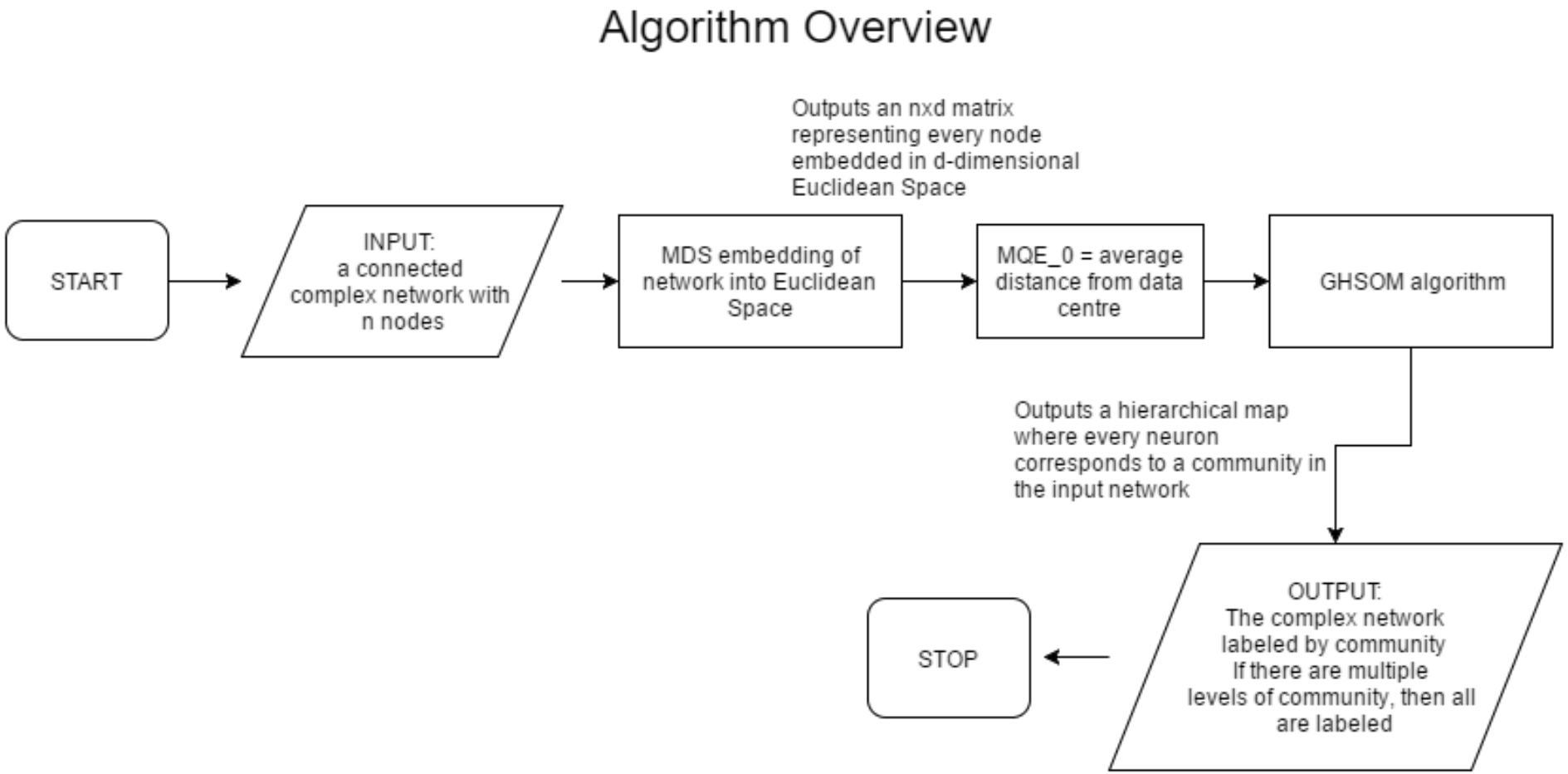
University of Birmingham,
Birmingham, UK
B15 2TT

27th February 2017

Recap

- Using variant of SOM to detect community structure in networks.
- Most biological networks contain multi-scale (hierarchical) community structure.
- The Growing Hierarchical Self Organizing Map (GHSOM) can grow maps of arbitrary shape and topology and select areas of the input space for a more fine-grain mapping.
- Use multi-dimensional scaling to embed nodes into Euclidean space for clustering, preserving shortest path distance.

Recap: Algorithm



Recap: Parameters

- η: learning rate
- w: initial weight range
- σ: neighbourhood function
- ϵ_{sg}: stop map growth parameter
- ϵ_{en}: expand parameter

Recap: GHSOM algorithm

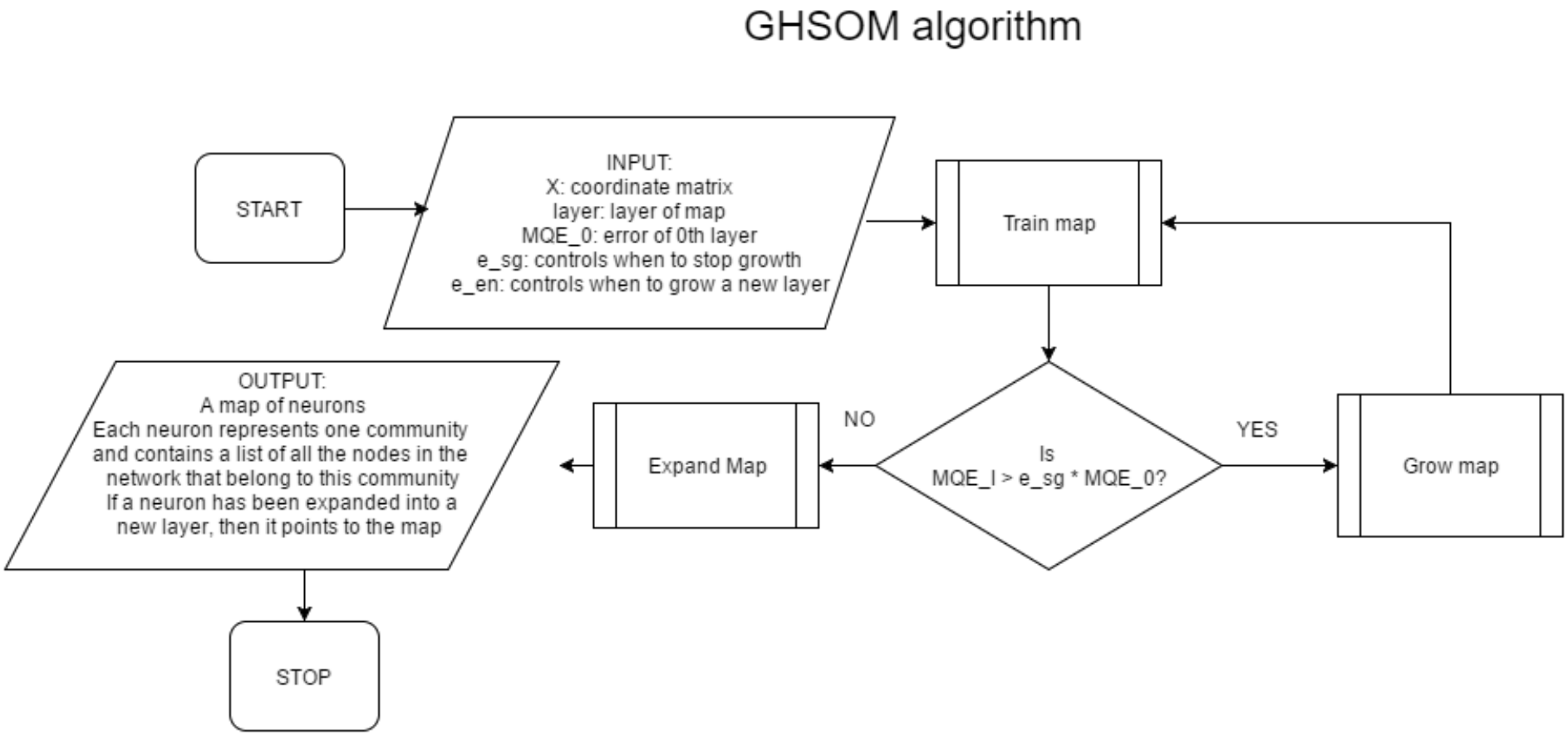


Figure: GHSOM algorithm first proposed in Dittenbach *et al.* [2000]

Results on synthetic benchmarks

- Evaluation on synthetic benchmark networks.
- Following example of Lancichinetti and Fortunato [2009] and Yang *et al.* [2013].
- 512 nodes.
- 4 macro communities for 128 nodes.
- Divided into 16 micro communities of 32 nodes.
- Fix number of connections to nodes in same micro community z_1 and connections to nodes in same macro community z_2 to 16.
- Vary the number of connections to nodes of other macro communities z_3.
- Record normalized mutual information (NMI) score.

Results on Synthetic Networks I

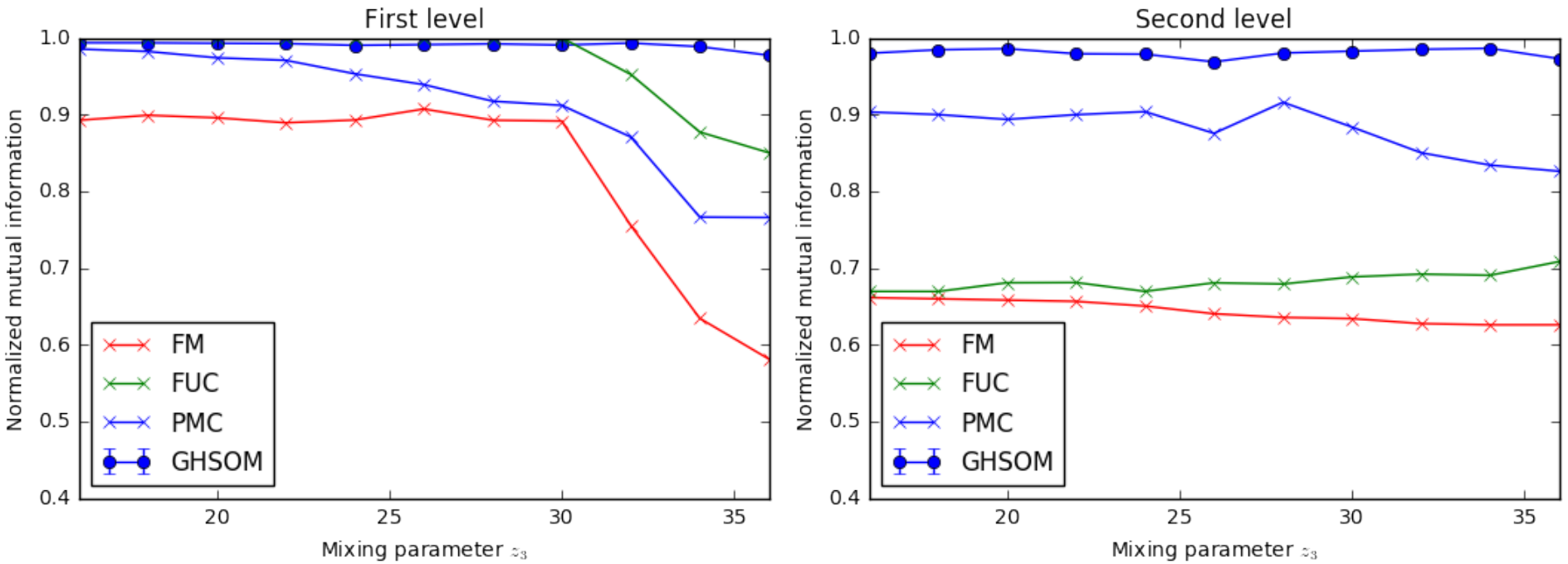


Figure: Plot of mixing parameter z_3 against NMI score for both levels of community. 100 networks were generated with each mixing parameter. Results extracted without permission from Yang *et al.* [2013] (fig 4).

Results on Synthetic Networks II

- FM: (Fast Modularity) – hierarchical agglomerative algorithm (Clauset *et al.* [2004])
- FUC: (Fast Unfolding of Communities) – heuristic method based on modularity optimization (Blondel *et al.* [2008]).
- PMC: (Probabilistically Mining Communities) – models community detection as a constrained quadratic optimization problem and solves using a random walk heuristic (Yang *et al.* [2013]).

Real world networks I

- What about real world networks?
- Limited to small real world benchmarks due to issues of scalability:
  - Karate
  - Dolphin
  - Polbooks
  - Football

Real world networks II

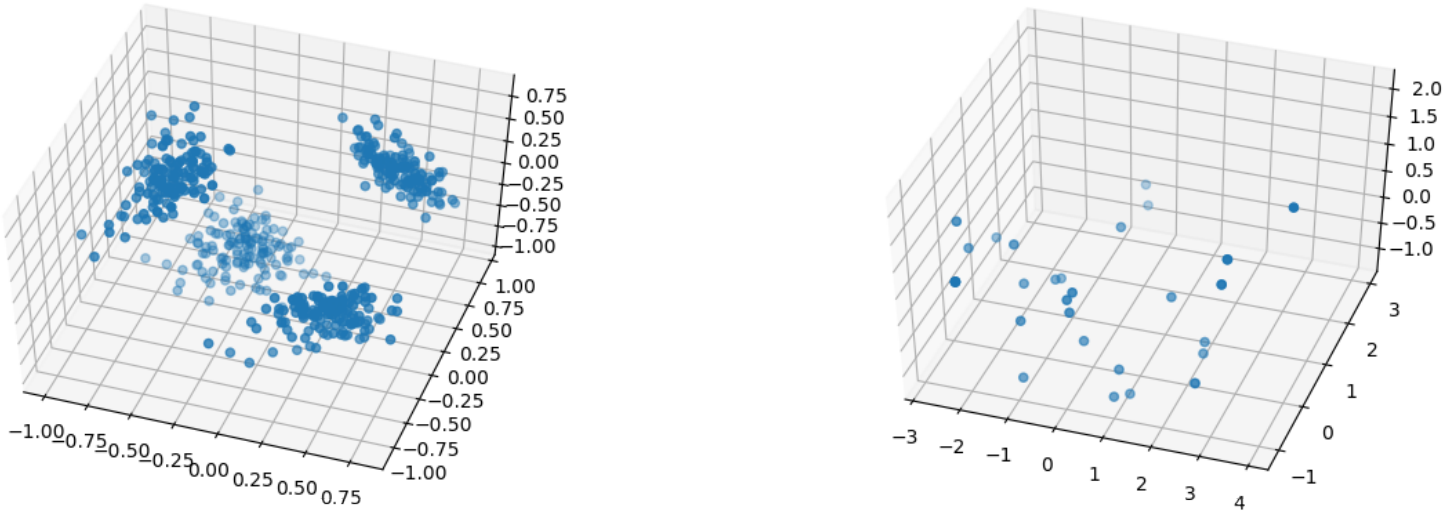


Figure: First three dimensions of example networks embedded using MDS.

- Parameter setting for the synthetic experiment was easier since the network was 'regular'.

Real world networks III

- Setting parameter settings for the real world networks is more challenging.
- A setting that was good for the karate network (with 2 labelled communities was not good for the football network (with 12 labelled communities.

Preliminary Results

Algorithm	**Network (NMI Score)**			
	Karate	Dolphin	Polbooks	Football
#comms	2	4	3	12
MCL	**1.000**	0.424	0.515	**0.935**
FM	0.693	0.509	0.531	0.757
FUC	0.587	**0.636**	**0.575**	0.855
PMC	0.837	0.620	0.574	0.887
GHSOM ($\epsilon_{sg} = 0.6$)	0.500	0.523	0.516	0.739
GHSOM ($\epsilon_{sg} = 0.8$)	0.733	0.575	0.547	0.528

Table: Table of NMI scores of GHSOM versus several algorithms in the literature. The best NMI score for each network is written in bold. Results for comparision algorithms are taken from Yang *et al.* [2013] (table 2) without permission. Markov Clustering algotithm (MCL) proposed by Van Dongen [2001].

Parameter Optimization I

Parameter	**Network**			
	Karate	Dolphin	Polbook	Football
η	0.0001	0.879	0.999	0.0587
w	0.943	0.0001	0.353	0.0663
σ	0.817	0.001	0.650	1.0
ϵ_{sg}	0.988	0.558	1.0	0.451
ϵ_{en}	1.0	0.3	0.393	0.3
#comms	2	4	3	12
#comms det.	2	4	2	11
NMI score	**1.0**	**0.640**	**0.688**	0.874

Table: Spearmint optimized parameter settings and NMI scores for real world networks (to 3 s.f.). Spearmint provided by Snoek *et al.* [2012].

Parameter Optimization II

- This looks bad.
- GHSOM does not seem very robust and is dependent upon parameter settings.
- But some good results, so there is potential.

Towards a more principled setting of parameters I

- Parameters:
  - η: learning rate
  - w: initial weight range
  - σ: neighbourhood function
  - ϵ_{sg}: stop map growth parameter
  - ϵ_{en}: expand parameter
- Fix every parameter except the stop growth parameter ϵ_{sg}.
- Derive equation for parameter setting ϵ_{sg} in terms of network density ρ.

Towards a more principled setting of parameters II

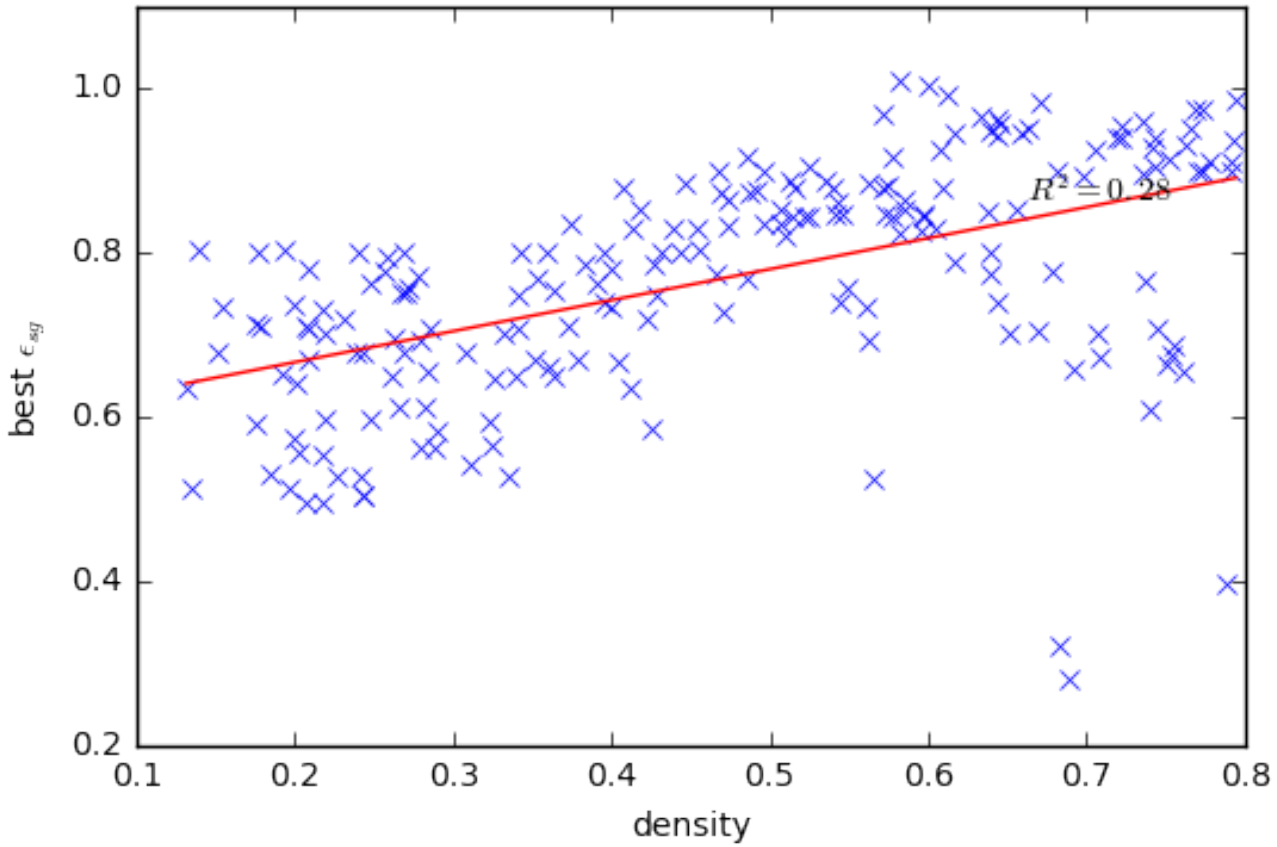


Figure: Plot of network density ρ vs. best setting for ϵ_{sg} found by 50 iterations of simulated annealing. 200 random networks of 64 nodes were generated with a random number of edges and communities.

Towards a more principled setting of parameters III

- Derived equation: $\epsilon_{sg} = 0.377746404462 * \rho + 0.590217653032$

Network	ρ	ϵ_{sg}	**Mean NMI Score**
Karate	0.139037433155	0.642738543492	0.499800072199
Dolphin	0.0840824960338	0.621979513587	0.493281211109
Polbooks	0.0807692307692	0.620727939546	0.518261769794
Football	0.0939740655988	0.625716018425	0.733839840555

Table: Derived ϵ_{sg} results. Each network was repeated 100 times.

A more realistic network density

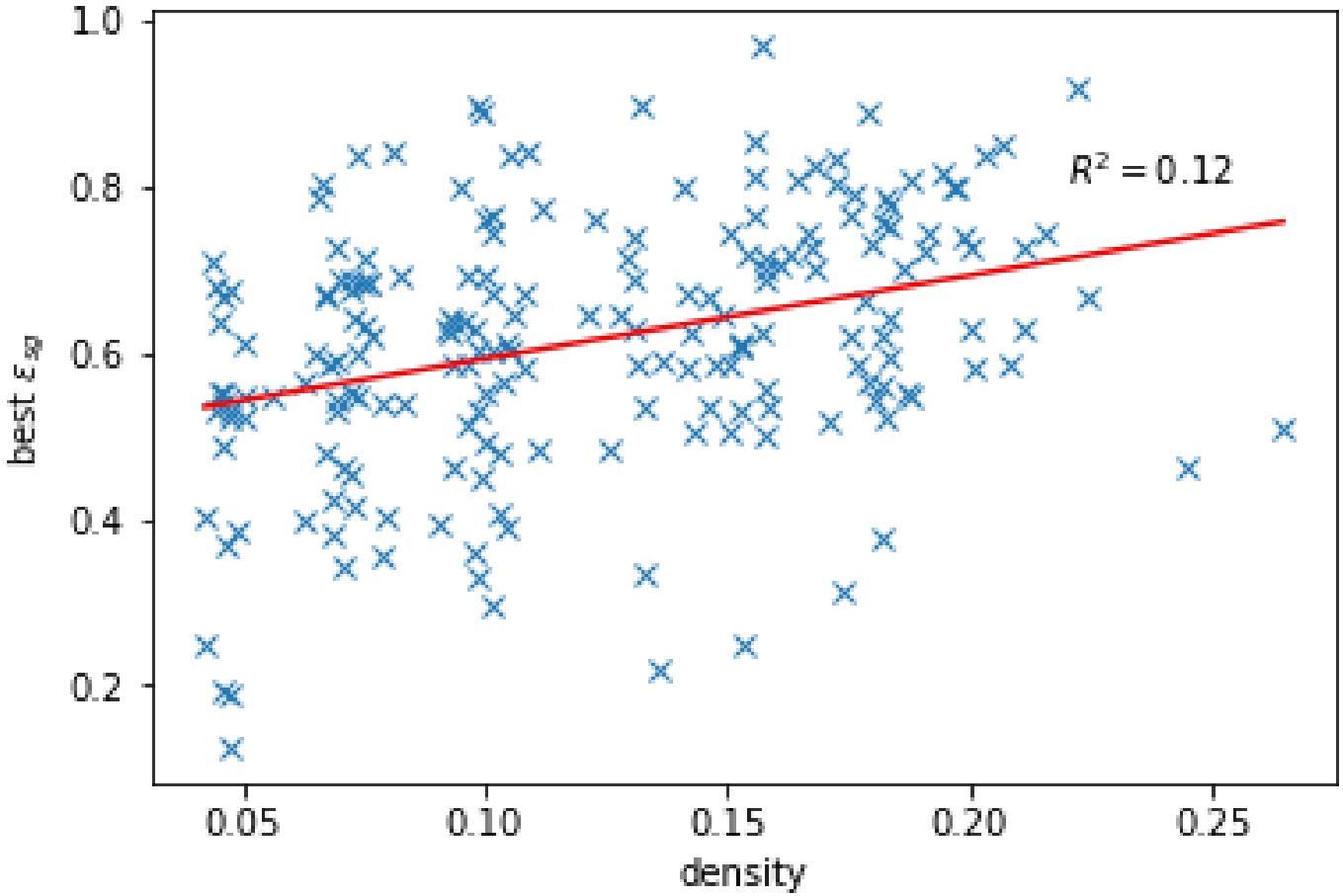


Figure: Plot of network density vs. best setting for ϵ_{sg} found by 50 iterations of simulated annealing. 200 random networks of 64 nodes were generated with a random number of edges and communities. Equation: $\epsilon_{sg} = 1.002 * \rho + 0.492$

(Some of a) Discussion

- Good when networks are regular.
- Very good results possible, even when mixing factor between communities is large.
- Parameter setting is challenging.
- Looking for a more principled approach.
- Poor scalability.
  - Experimentation is time-consuming.
  - Large real-world networks are infeasible.
  - Could be efficiently parallelized.

What's next: Topology Preservation

- The main advantage of SOM is topology preservation.
- Similar features are mapped close together.
- Is this true for community detection?
- Will communities that are mapped together by GHSOM have similar functions?
- Experiment on small unlabelled biological network and compare GO terms.

References I

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

Michael Dittenbach, Dieter Merkl, and Andreas Rauber. The growing hierarchical self-organizing map. In *IJCNN (6)*, pages 15–19, 2000.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

References II

Stijn Marinus Van Dongen. *Graph clustering by flow simulation*. PhD thesis, 2001.

Bo Yang, Jin Di, Jiming Liu, and Dayou Liu. Hierarchical community detection with applications to real-world network analysis. *Data & Knowledge Engineering*, 83:20–38, 2013.