Growing Hierarchical Self Organizing Maps for Community Detection

David McDonald Shan He

10th April 2017

Recap

- Using Growing Hierarchical Self Organising Maps to detect communities at multiple scales in complex networks.
- Achieved very good NMI scores on synthetic hierarchical benchmarks.
- Was capable of good NMI scores on real world benchmarks, but finding a principled methodology of setting parameters was challenging.

Results on Synthetic Networks

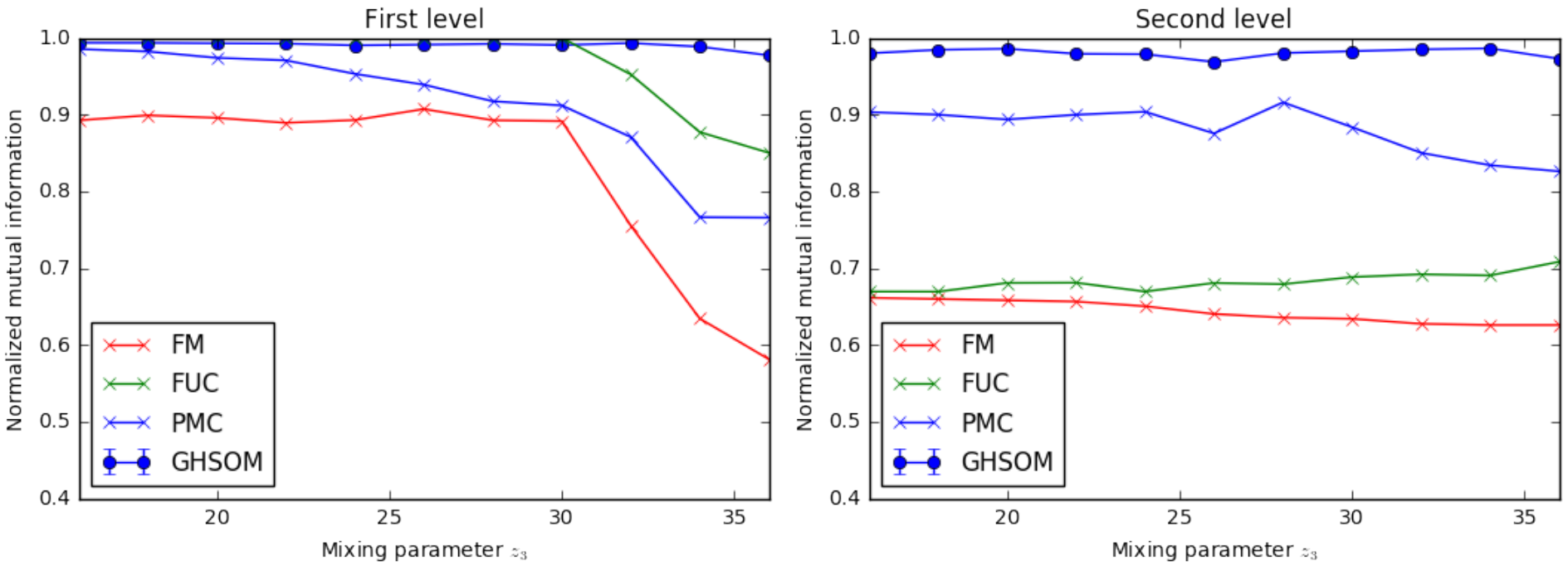

Figure: Plot of mixing parameter z_3 against NMI score for both levels of community. 100 networks were generated with each mixing parameter. Results extracted without permission from Yang *et al.* [2013] (fig 4).

Results on Real World Networks

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 comparison algorithms are taken from Yang *et al.* [2013] (table 2) without permission.

Bayesian Optimisation

Parameter	**Network (NMI Score)**			
	Karate	Dolphin	Polbook	Football
#comms	2	4	3	12
η	0.0001	0.879	0.999	0.0587
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 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].

Current Research: Topological Functional Similarity Neighbouring Communities

Do neighbouring communities cooperate?

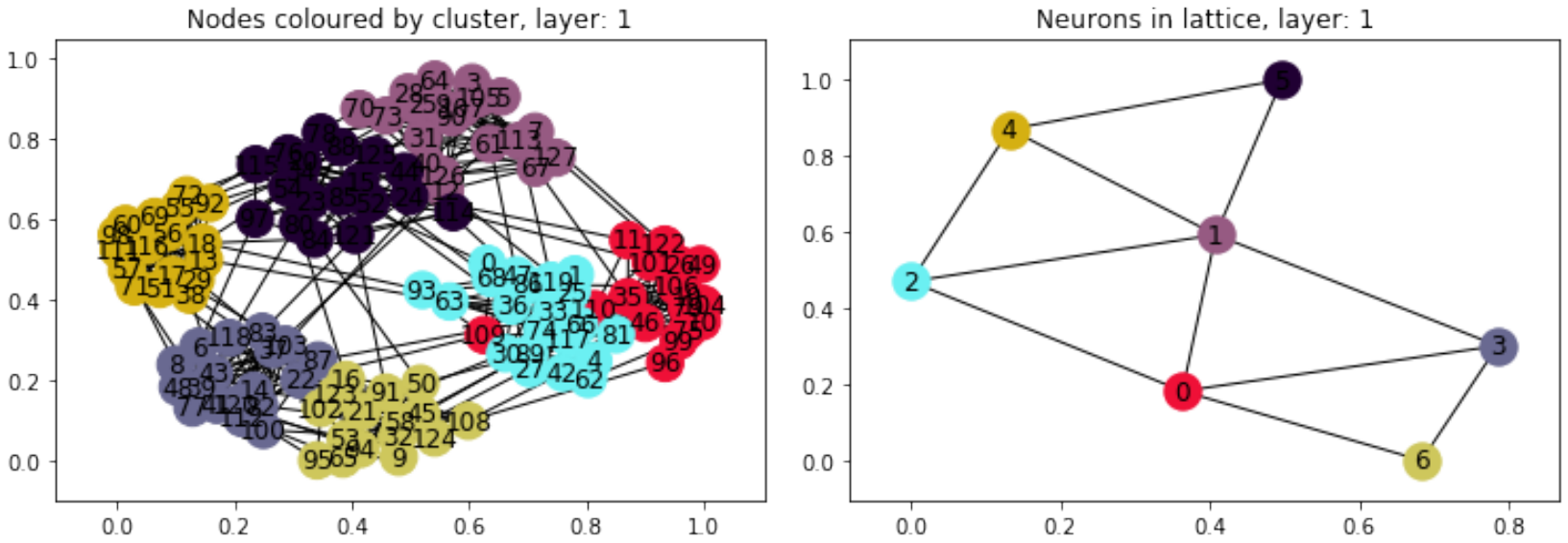

Figure: Visualisation of simple generated network and resulting map. Nodes are coloured in the network by the neuron in the map that they are assigned to.

A Little Background: The Gene Ontology (GO)

- A controlled vocabulary to share information about all genes across all *eukaryotes* (organisms with cells containing a nucleus) (Ashburner *et al.* [2000]).
- Represented as three directed acyclic graphs (DAGs) corresponding to the three ontologies: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).
- Nodes on the graph are GO terms and edges are relations.
- Genes are annotated with GO terms using annotation databases.
- All annotations obey the true path rule: if a gene is annotated with a term then it is automatically annotated with all of that term's ancestors.

	GOID	TERM	ONTOLOGY
1	GO:0007323	peptide pheromone maturation	BP
2	GO:0018342	protein prenylation	BP
3	GO:0018344	protein geranylgeranylation	BP
4	GO:0018343	protein farnesylation	BP
5	GO:0005953	CAAX-protein geranylgeranyltransferase complex	CC
6	GO:0005965	protein farnesyltransferase complex	CC
7	GO:0004661	protein geranylgeranyltransferase activity	MF
8	GO:0016740	transferase activity	MF
9	GO:0004660	protein farnesyltransferase activity	MF
10	GO:0008318	protein prenyltransferase activity	MF
11	GO:0004659	prenyltransferase activity	MF
12	GO:0004662	CAAX-protein geranylgeranyltransferase activity	MF
13	GO:0004660	protein farnesyltransferase activity	MF
14	GO:0004662	CAAX-protein geranylgeranyltransferase activity	MF

Figure: All GO terms annotated to the gene with ORF identifier YKL019W in the org.Sc.sgd.db annotation database. YKL019W is automatically annotated with the parents of all these terms due to the true path rule of GO.

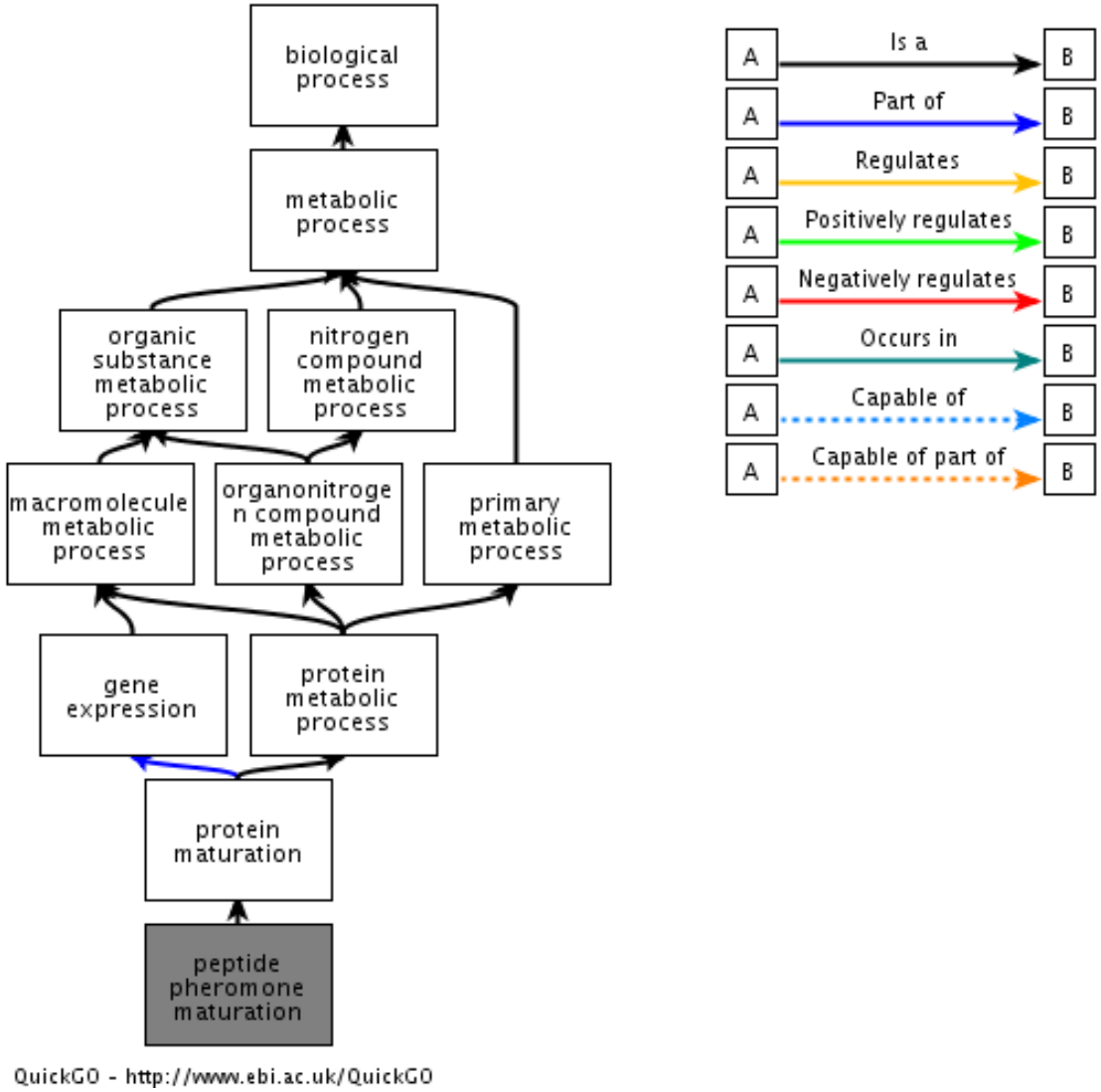

Figure: All ancestors of the GO term GO:0007313 *peptide pheromone maturation* in the BP ontology.

Experiments on Saccharomyces Cerevisiae

- Experimented on two *Saccharomyces Cerevisiae* (budding yeast) co-expression networks.
- Found largest fully connected component in each network and embedded using MDS, based on shortest path between each pair of genes in the network.

Network name	**Number of Nodes**	**Number of Edges**
Uetz Screen	263	292
Y2H Union	1647	2682

Table: Topology information for the two Saccharomyces Cerevisiae networks. Datasets available from uet; uni.

Method

- Use GHSOM to partition network into set of communities.
- For each community, determine the set of enriched GO terms using a 2x2 contingency table and Fisher's exact test.
- Select terms based on p-value 0.05.

	sig	notSig
anno	9	6
notAnno	33	202

Figure: An example contingency table for the go term: GO:0006914 *autophagy* (p-value=0.00011102437580897 3).

Similarity Measures

Two similarity measures used so far. Both very prominent in the literature.

- Resnik and others [1999]
  - Originally used on words.
  - Assigns measure of *information content* based on number of offspring.
  - Lower terms in the DAG contain more information.
  - The similarity of two terms is the greatest information content of their common ancestors.
  - "Resnik's measure correlates well with gene expression" (Sevilla *et al.* [2005]).
- Wang *et al.* [2007]
  - Based on topology of DAG.
  - Weightings are assigned to each relation.
  - Recursively assign semantic value to each term in the DAG induced from a given term, by using edge weights.
  - Semantic similarity of two terms proportional to the number of terms in both DAGs.

Results

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6	Com 7	Com 8
Com 1	1	0.472	0.415	0.443	0.444	0.386	0.473	0.317
Com 2	0.472	1	0.62	0.53	0.656	0.379	0.401	0.485
Com 3	0.415	0.62	1	0.55	0.566	0.442	0.423	0.512
Com 4	0.443	0.53	0.55	1	0.484	0.479	0.383	0.356
Com 5	0.444	0.656	0.566	0.484	1	0.43	0.43	0.534
Com 6	0.386	0.379	0.442	0.479	0.43	1	0.43	0.287
Com 7	0.473	0.401	0.423	0.383	0.43	0.43	1	0.357
Com 8	0.317	0.485	0.512	0.356	0.534	0.287	0.357	1

	Com 1	Com 2	Com 3	Com 4	Com 5	Com 6	Com 7	Com 8
Com 1	0	3	4	1	5	7	1	4
Com 2	3	0	1	4	2	4	4	1
Com 3	4	1	0	5	3	5	5	2
Com 4	1	4	5	0	6	8	2	5
Com 5	5	2	3	6	0	6	6	3
Com 6	7	4	5	8	6	0	8	5
Com 7	1	4	5	2	6	8	0	5
Com 8	4	1	2	5	3	5	5	0

Figure: Example results obtained with the Wang *et al.* [2007] similarity measurement. 51 communities were found by GHSOM. *Left:* similarities of first 8 communities. *Right:* shortest path length of first 8 communities on map.

One Possible Explanation

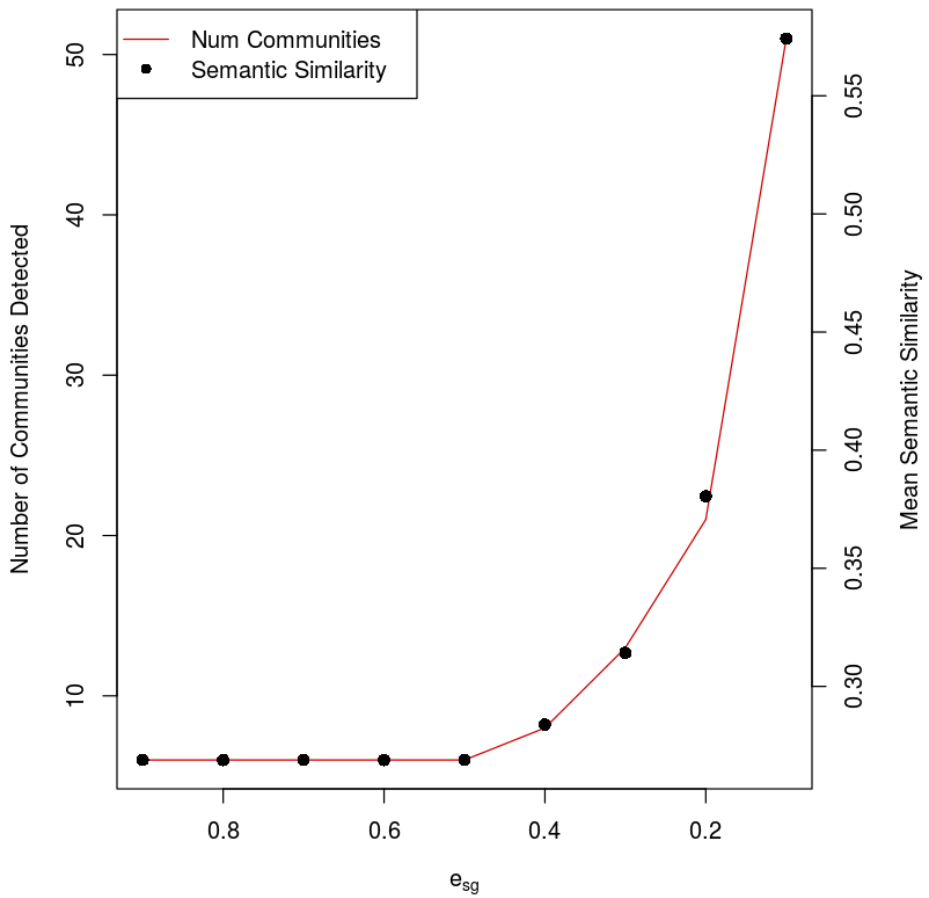

Figure: Plot of e_{sg} against the number of communities detected in the Uetz screen network and the mean functional similarities of genes in the same community.

Plans to Finish

- Search for much smaller communities – that maximise the functional similarities of clusters.
  - Running right now...
- Try another type of network: Social networks
  - I have partitioned the Florentine families network and just need ot analyse the results against the ground truths found by other papers.
- Compare with results of hierarchical clustering.
- Can embedding based on functional similarity rather than shortest path distance produce better results?

References I

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Philip Resnik et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.(JAIR)*, 11:95–130, 1999.

Jose L Sevilla, Victor Segura, Adam Podhorski, Elizabeth Guruceaga, Jose M Mato, Luis A Martinez-Cruz, Fernando J Corrales, and Angel Rubio. Correlation between gene expression and go semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4):330–338, 2005.

References II

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.

Uetz screen dataset. `http://interactome.dfci.harvard.edu/S_cerevisiae/download/Uetz_screen.txt`.

Y2h union dataset. `http://interactome.dfci.harvard.edu/S_cerevisiae/download/Y2H_union.txt`.

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

James Z Wang, Zhidian Du, Rapeeporn Payattakool, S Yu Philip, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

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.