A Nonparametric Significance Test for Sampled Networks

Andrew Elliott

University of Oxford

September 16, 2013

Motivation

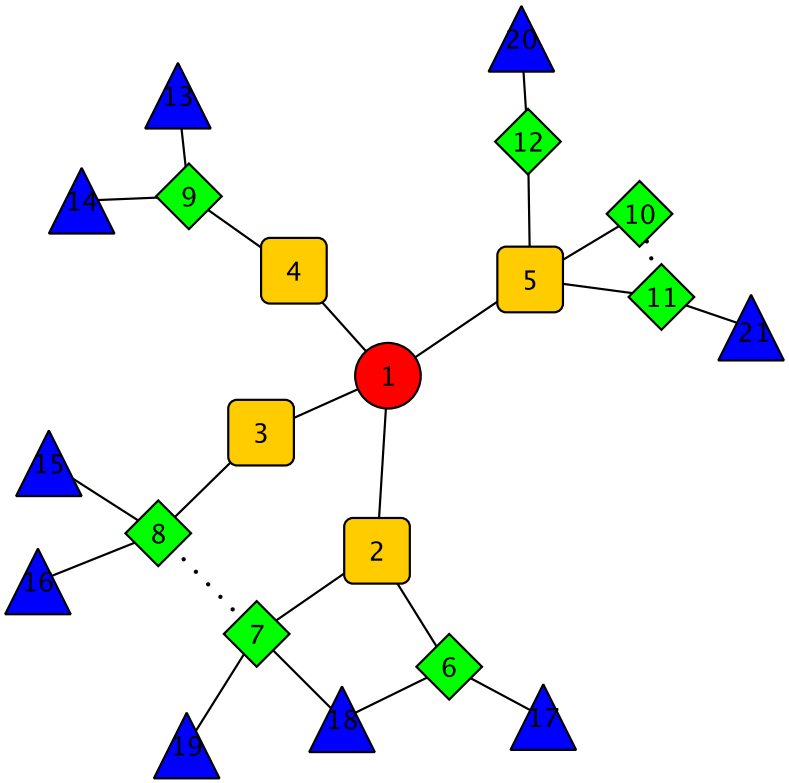
- Most of the PPI (protein-protein interaction) data is global
- We want to construct a PPI (protein-protein interaction) network related to Parkinson's disease (PD)
- There are many different ways to construct such a network
- Choosing between them is difficult, an appropriate null model can solve this problem
- Many standard null models do not account for the sampling technique

Subsampling in Biology

- In the last ten years there has been an explosion in high throughput biological interaction data
- However this data is (generally) not specific to a particular problem
- One method to generate a specific subnetwork is to subsample
- The question is: When is such a subnetwork informative?
- Perhaps when a selection of network statistics are different from random?

Subsampling in Biology (Part II)

- Most methods for generating biologically relevant subnetworks typically involve:
  - A seed list
  - A rule for construction



So what are Protein-Protein Interaction Networks?

- Proteins are the building blocks of cells
- A PPI is as a pair of proteins that molecularly dock
- Database of PPIs contain high levels of false positives and negatives
- Importantly there is no good general generative model of PPI networks

Seed lists

We assemble 2 seed lists, each based on a different data source:

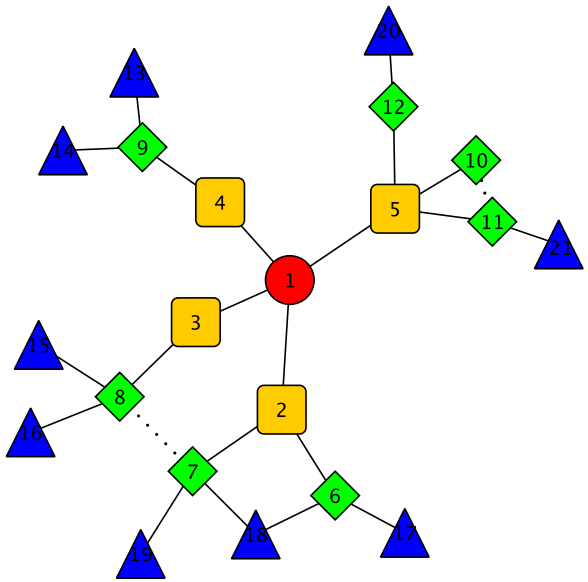
Expression Data Conn et al. measured the differential expression of human cells before and after treatment with the toxin MPP+ (a known PD disease model) [1]

Disease Genes Parkinson's related disease genes from OMIM database [2]

Each seed list only includes seeds that are in the largest connected component of the PPI network

1. Conn,K.J. *et al*. (2003) cDNA microarray analysis of changes in gene expression associated with MPP+ toxicity in SH-SY5Y cells. *Neurochem Res*, **28** 1873–1881.
2. Hamosh,A. *et al*. (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, **33 (suppl 1)** D514–D517.

Sampling Techniques: Snowball Sampling



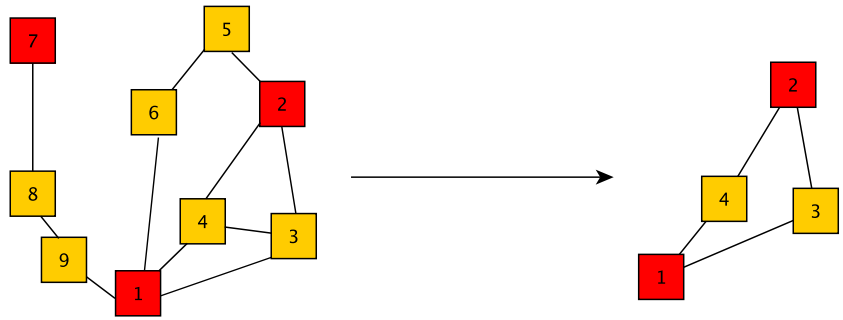
1. Snowball Sampling: a **'guilt by association'** approach. Used in Biology [1] and Sociology [2,3]

   Limited snowball sampling to 1 and 2-hop sampling

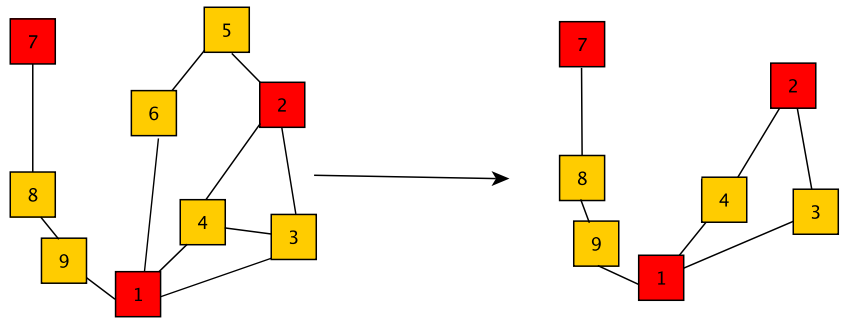
1. Martin,A. *et al.* (2010) BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics*, **11** 91+.
2. Frank,O. (1977) Survey sampling in graphs. *J Stat Plan Infer*, **1** 235–264.
3. Bernard,H.R. *et al.* (2010) Counting hard-to-count populations: the network scale-up method for public health. *Sex Transm Infect*, **86** ii11-ii15

Sampling Techniques

2. Path $\leq$ 'k': Short routes between seed nodes are important. Used in the Genes2Networks application [1]. Limited to k = 2, 3, 4



3. 'Shortest paths': shortest paths between seed proteins are important



1. Berger,S. *et al*. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8** 372.

Candidate Null Models

- We have 12 subnetworks supposedly related to PD
- We need a null model to compute the significance of network features
- Potential Null Models:
  1. Basic ER graphs
  2. Configuration Model
  3. Synthetic PPI networks

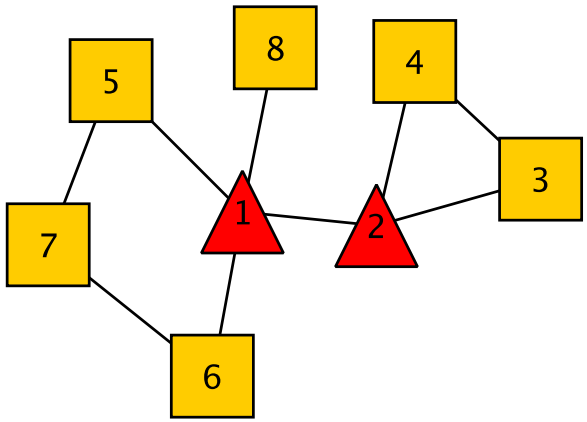
Our Proposed Null Model

- Resampling the original network gives us a **credible null model**
  - Has all of the features of a PPI network
  - Replicates the sampling process

- However, we are interested in the significance of network features not of seed lists
  - Many different seed lists will give the same resultant network
  - Thus a null model based purely on the seed list can be problematic
  - We must control for the seed list construction

Controlling for Seed List Construction

Minimum Seed List (MSL) The set of the smallest subsets of the original seed list that generate the same subnetwork



Red Triangles = Seed Nodes

- 1-hop Snowball Sampling: The MSL is the $\{1,2\}$
- 2-hop Snowball Sampling: The MSL is $\{1\}$

Computing the Minimum Seed List

- Finding the MSL is computationally expensive
- For Snowball Sampling it reduces to a NP complete problem
- In other sampling techniques reduces to intelligently checking combinations of nodes that can be removed individually

The Null Model

1. Find the minimum seed list.
2. Construct an ensemble of random seedlists of the same degree sequence as the original seed list
3. Construct an ensemble of subnetworks by subsampling the original network using the seedlist ensemble
4. Use the distribution of the statistic of interest on the ensemble of subnetworks

Analytics (1)

Possible in simple cases.

$$E(X|S{=}s) = |V| - \sum_{\substack{J \subset V \\ |J|=1}} h(J, s)$$

X – #Nodes in n-hop Snowball sampling
S – Length of the seedlist
V – Set of nodes in the original graph
$B(M)$ – Set of nodes sampled in n-hop snowball sampling with set of seeds M

$$h(M, s) = \prod_{i=0}^{s-1} \frac{|V| - |B(M)| - i}{|V| - i}$$

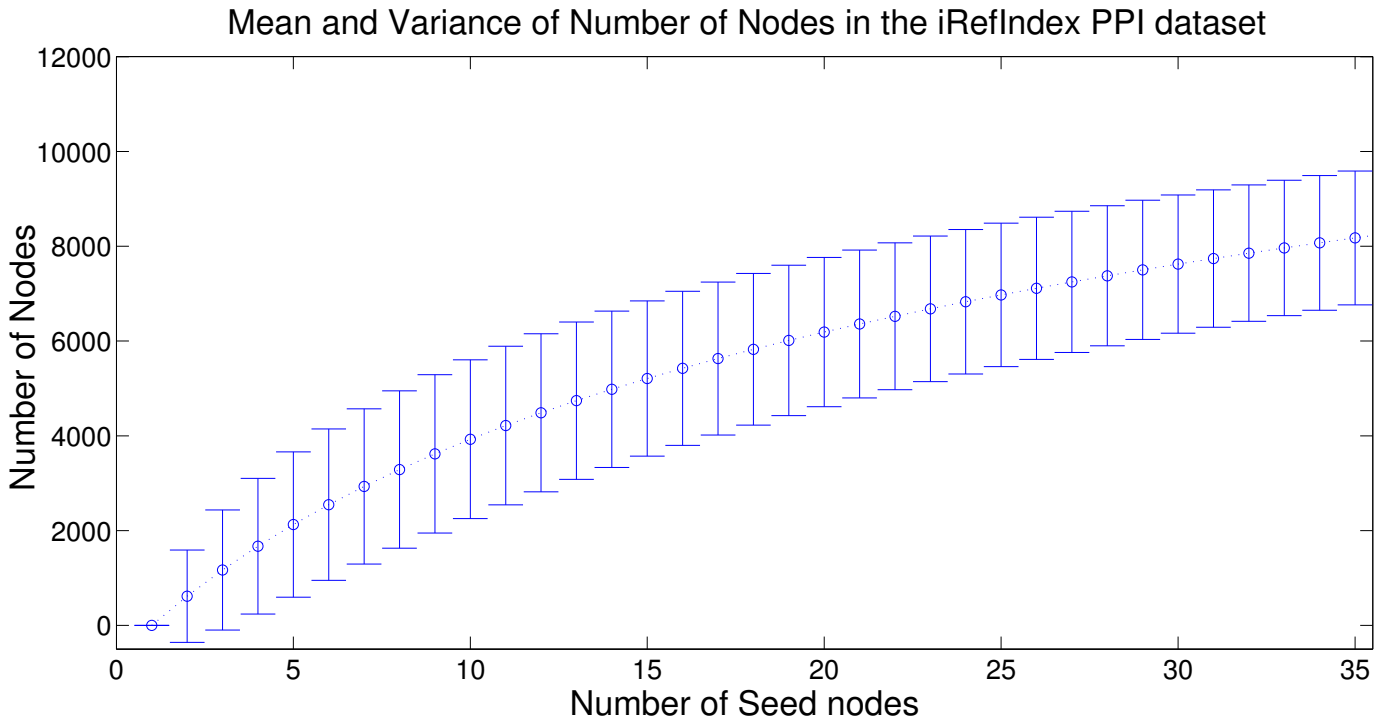
Analytics (2)

$$\mathrm{Var}(X|S{=}s) = (1 - \sum_{\substack{J \subset V \\ |J|=1}} h(J,s)) \sum_{\substack{J \subset V \\ |J|=1}} h(J,s) + 2 \sum_{\substack{L \subset V \\ |L|=2}} h(L,s)$$

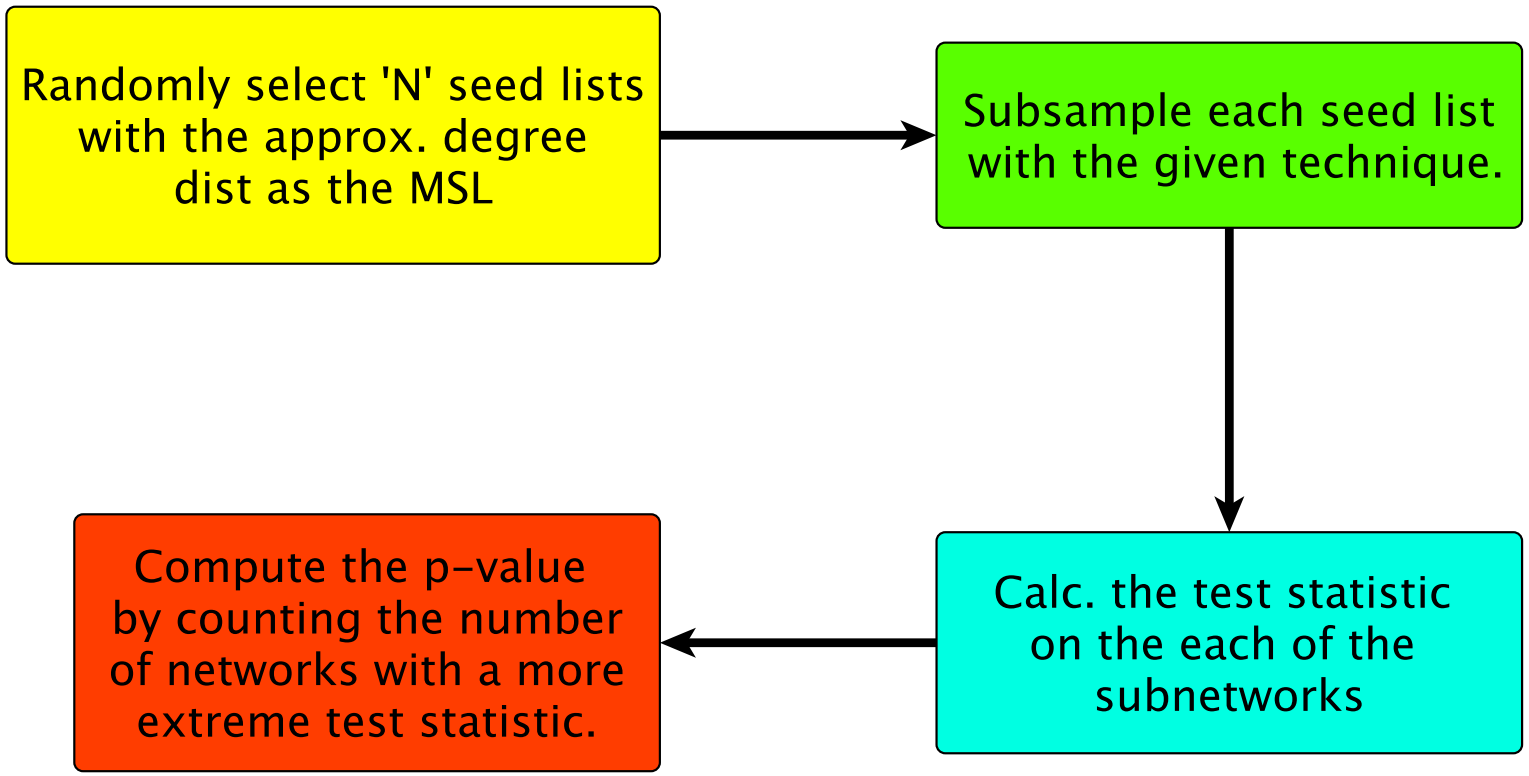
X – #Nodes in n-hop Snowball sampling
S – Length of the seedlist
V – Set of nodes in the original graph
$B(M)$ – Set of nodes sampled in n-hop snowball sampling with set of seeds M

$$h(M,s) = \prod_{i=0}^{s-1} \frac{|V| - |B(M)| - i}{|V| - i}$$

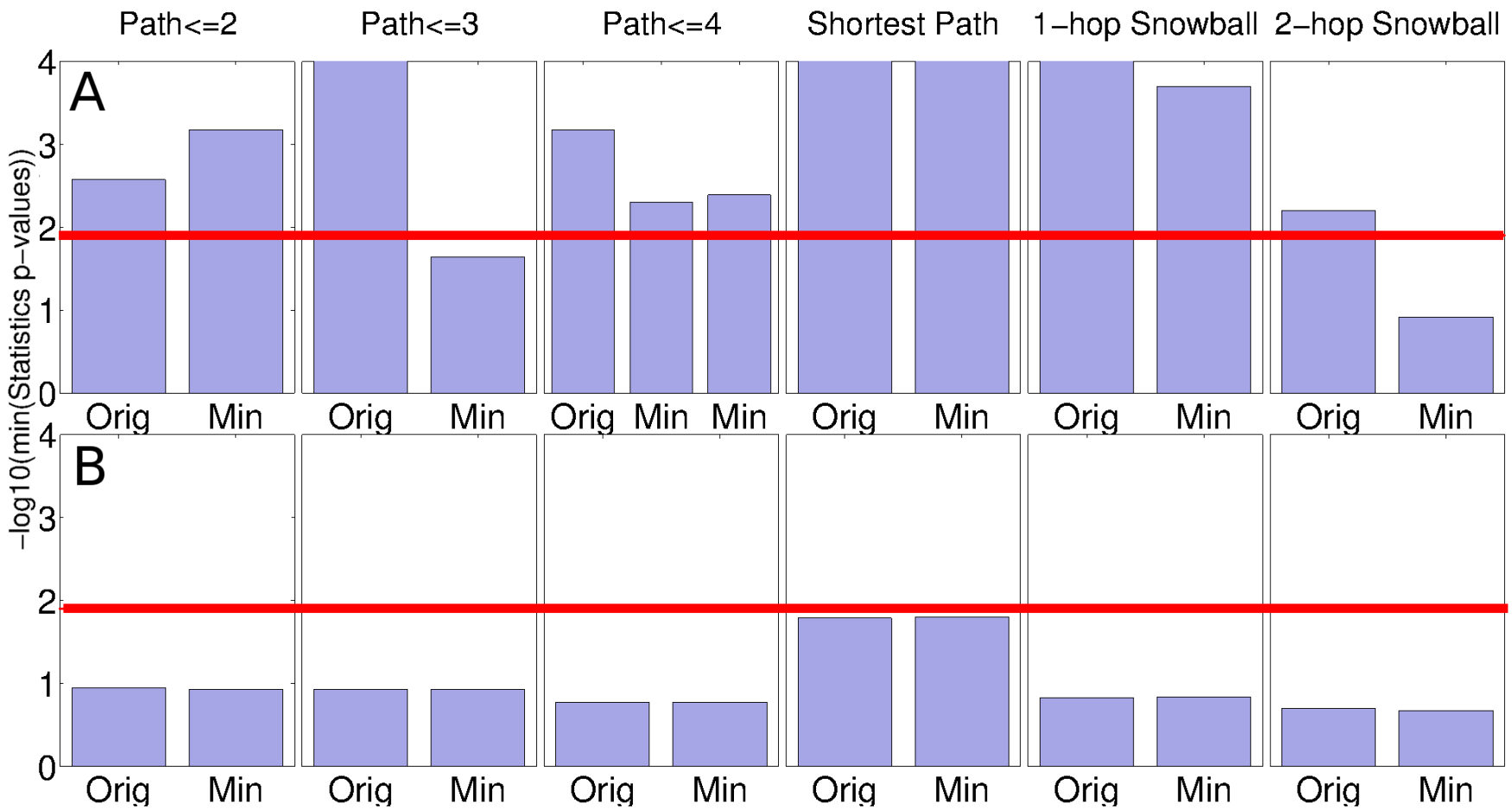
Exact Results



Monte Carlo Approach



Results: Empirical Seed Lists

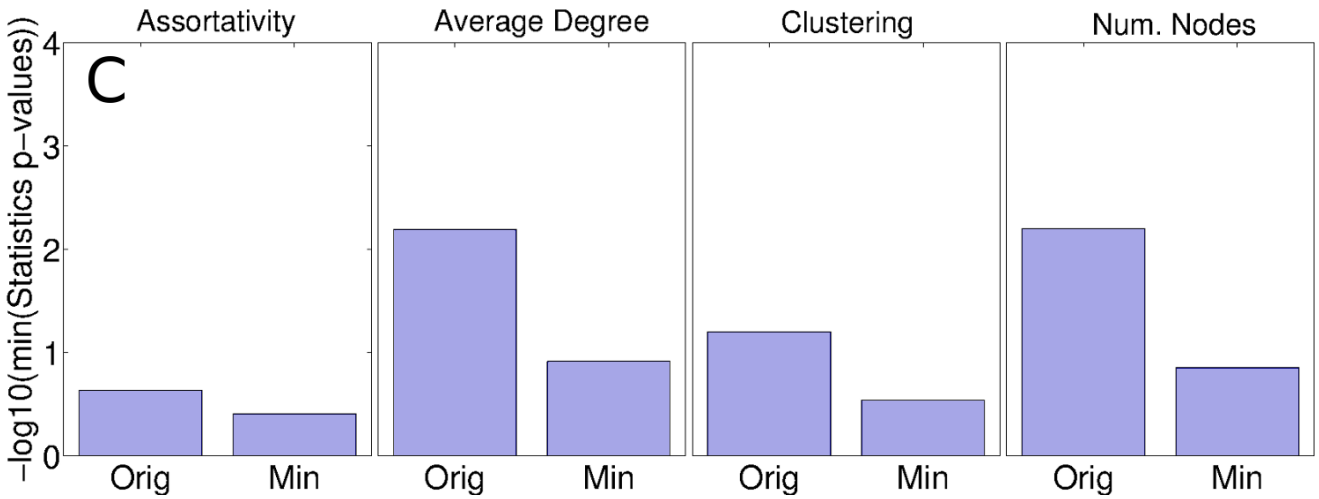


A OMIM seed list

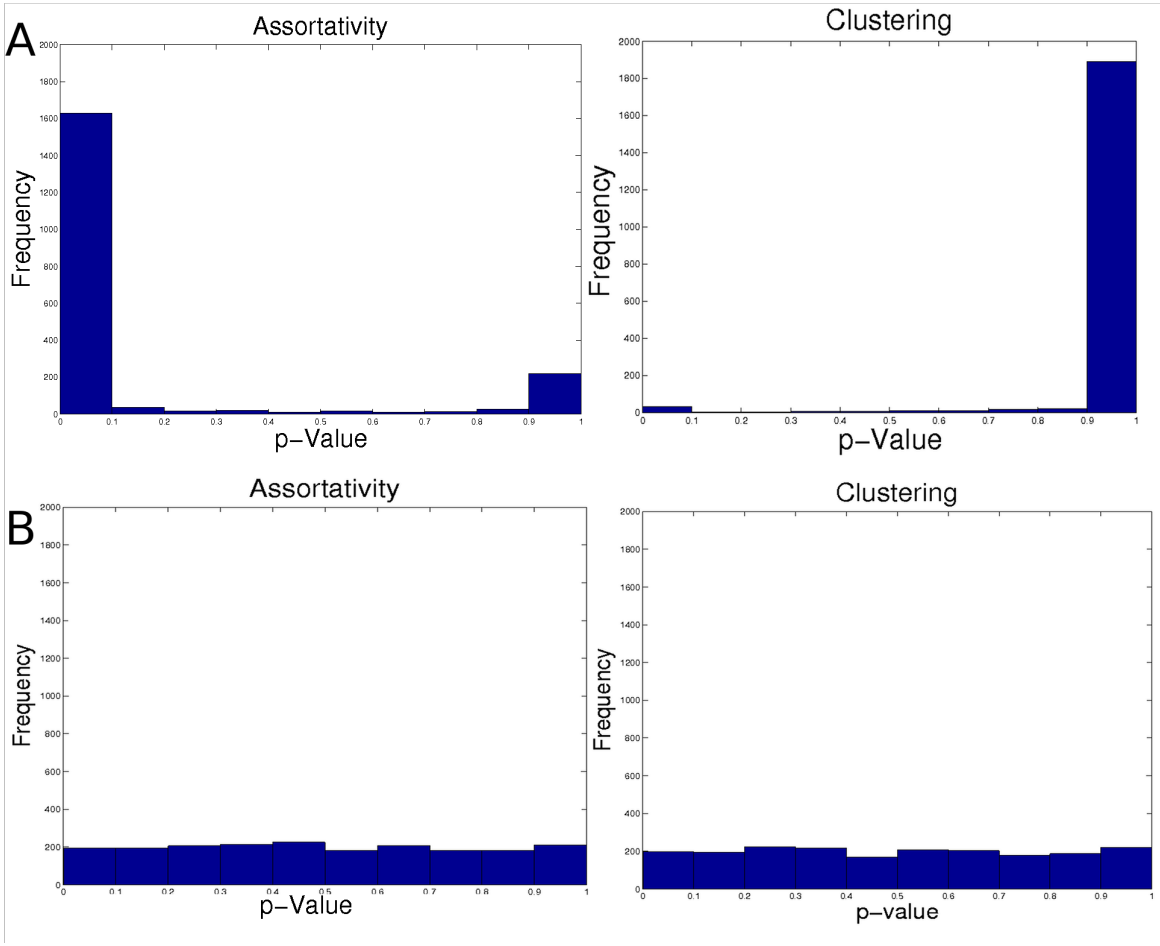
B Expression seed list.

Results: Individual Statistics

- MSL influence on significance is not limited to one statistic.
- OMIM 2-hop snowball sampled network:



Comparison with Configuration Model



- We assess the uniformity of the distribution using a χ^2 test.
- Results (p-values):
  - Configuration Model: numerically equal to 0
  - New Model: 0.42 in Assortativity and 0.1 in Clustering Coefficient

Conclusions

- We provide a null model for sampled networks.
- We use this null model to discriminate between subnetworks generated using different seed lists and sampling techniques.
- The subnetworks based on expression data do not differ from random.
- Subnetworks created from an OMIM MSL differ significantly from random, under shortest path and 1-hop snowball sampling.
- Redundant seed nodes have to be removed before comparison.
- The configuration model is not suitable for this task.

Any Questions?

Acknowledgements
Project supervised by:

- Dr Felix Reed-Tsochas (CABDyN, Said Business School, Oxford),
- Prof. Gesine Reinert (Dept. of Statistics, University of Oxford),
- Dr Elizabeth Leicht (CABDyN, Said Business School, Oxford)
- Dr Alan Whitmore (e-Therapeutics plc).