A parameterized approach to the analysis of Otitis Media data

A Collaborative project between the PCRU and
Menzies School of Health

Peter Shaw, Faisal N. Abu-Khzam, Robyn Marsh, Heidi Smith-Vaughan

Data: 2017/08/22

Charles Darwin University

Overview

1. Introduction
2. Motivation
3. Research Direction
4. Study Population and Design
5. Analysis Tools
6. Computing a Network Model
7. Results

Introduction

Motivation

Motivation

- Otitis Media (Ear Infect) is a major health issue in NT AU.
- 30% of Children are deaf. 99.7% of Indigenous inmates are deaf [5].
- Currently vaccines and antibiotics not sufficiently effective
- Multi-pathogen disease. See `http://www.ncbi.nlm.nih.gov/pubmed/23523773`
- PCA and GLM limited to $1 : many$ interactions
- Clinical researched requested Network Model analysis

Research Direction

Research Direction

PCRU has been developing a suit of tools for multivariate data analysis that utilizes FPT (Clique Centric) Algorithms

1. Fixed Parameter Tractable (FPT) $T(n, k) = a^k + O(n^c)$
2. Microbiological DNA, RNA, Protein-Protein interactions [2, 7]
3. Ecological Data Silos Northern Seas Ocean study [2]
4. **Currently focusing on Clinical Data**
   - Type I Diabetes
   - Otitis Media (OM)
   - ALRI [4]

Study Population and Design

Study Population and Design

- Random collection trial between 1996-2001,2001-2004
- swabs were collected from two months to 2 years of age
- 740 nasopharyngeal swabs (these are sequenced)
- to produce 29 variables
- Correlations were computing between all pairs of variables
- a threshold of 0.2 was used
- Key variates are Otitis Media (OM) and ALRI.

Unclustered Variables Hard to Understand

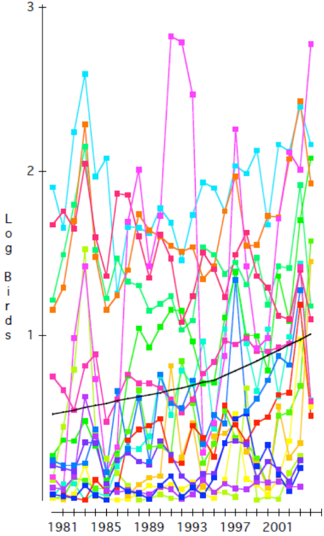

Figure: Unclustered variables are NP-Hard to unscramble.[2]

Analysis Tools

Clique Centric Pipeline

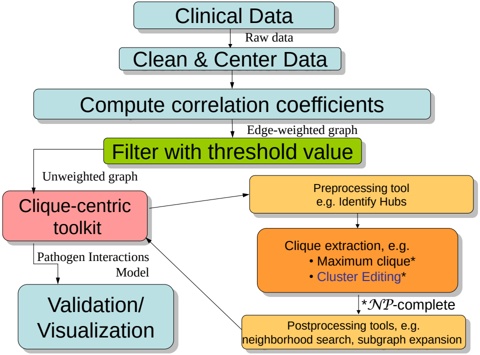

Figure: The role of Cluster Editing in the clinical analysis pipeline. This figure and methodology are an adaptation of work by Langston et al. [3]

Don’t compound the error

- Clinical, psychological and biological measurements can be fuzzy
- Many unsupervised clustering problems are NP-Hard
- FPT offers a way to compute exact solutions
- Why use approximations when you can calculate the exact solution

Computing a Network Model

CLUSTER EDITING Problem

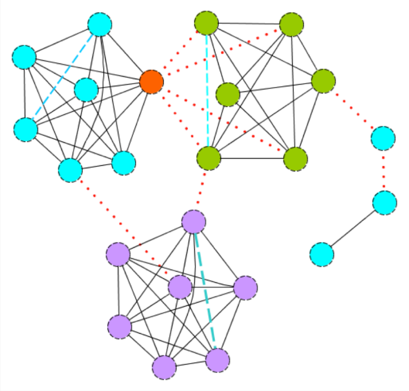

Figure: Added Edges colored blue. Deleted Edges Colored Red

- ◎ Missing edges indicate False-negative correlations

  Perhaps the relationship is not linear or effected by outliers
- ◎ Added edges indicate False-Positive correlations

Whats Novel

Whats new about how we analyise this Clinical Data.

Otitis Media (OM), Acute Lower Respitory Infections (ALRI) analysis and Type I Diabetes

- Heterogeneous Data
- Categorical, Continuous, Logical Data means
- New Measures of correlation
  - Spearmans Rank
  - Polychoric, Tetrachoric
  - Bi-Serial

Higher levels of Noise

- Lower thresholds for significant correlation
- could use *eigen values* but also has to make sense in the domain
- This means more Noise
- Faster Algorithms needed

Aggresize Parameterization

- Multi-parameter Cluster Editing

  $T(n, (k, a, d)) = \alpha^k + f(a) + f(d) + (n^c)$
- Bound the number of false +ve and −ve correlations that can be attributed to a single Variable.
- We need different values to deal with hubs.
- New (Hybrid) Graph Structure
- New algorithms faster and also models the noise better.

Results

Clinical

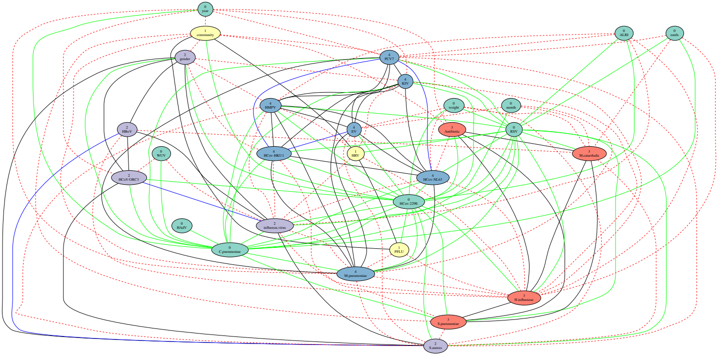

Figure: Pathway analysis of nasal pathogens reveals four distinct clusters using clique as a structural measure. (The diagram was produced using Graphvis by first weighting and coloring the graph based on the clusters found. [1]).

Clinical

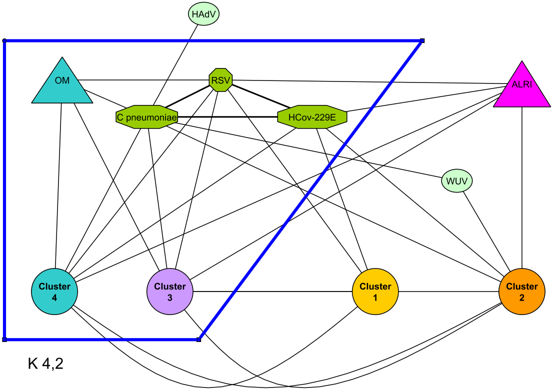

Figure: Pathway analysis of nasal pathogens revels four distinct clusters using clique as a structural measure.

Cluster 4

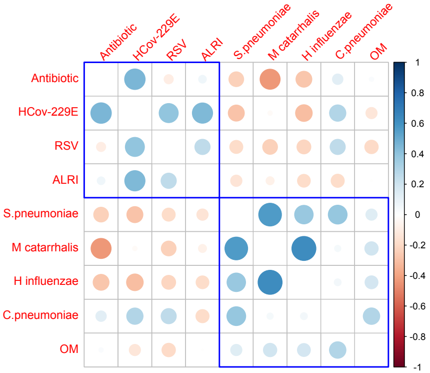

Figure: A heatmap of Cluster 4 with secondary hysterical clustering (produced using R:package corrplot [6])

With regard to the vaccination PCV7 we can see that this is

Cluster 3

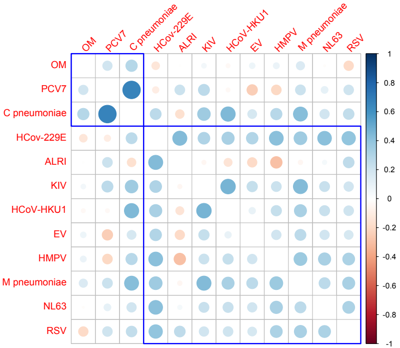

Figure: A heatmap of Cluster 3 with secondary hysterical clustering (produced using R:package corrplot)

Cluster 1

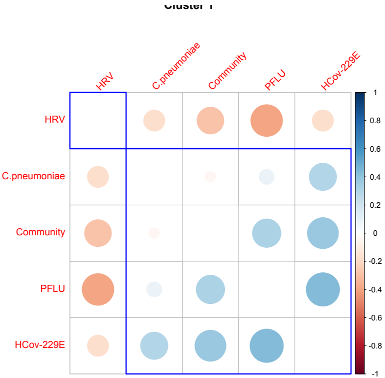

Figure: A heatmap of Cluster 1 with secondary hysterical clustering (produced using R:package corrplot [6])

Execution time

- Much faster >> 600 times
- Hybrid Data Structure
- Aggressive Multi-Parameter
- Hubs

Conclusion

Future Work

- Use $t(X)$ to cluster patient with similar symptoms
- Data Linkage data sets $50k - 100k$
- Spectral pre-processing
- Need good heuristics for these large data sets

Upcoming Publications

* **(1)** Does the Development of Hypotheses Stifle the Development of Knowledge? The Role of Cognitive Biases and the Benefits of Data Mining.

Simon Moss and Peter Shaw

* **(2)** Effective Use of Multi-parameterized Correlation Clustering in Mining Nasopharyngeal Carriage and Disease data from Young Children
Peter Shaw, Faisal N. Abu-Khzam, Robyn Marsh, Heidi Smith-Vaughan (2017)

* **(3)** Cluster Editing with Vertex Splitting
Faisal N. Abu-Khzam, Serge Gaspers, Alexis Shaw Peter Shaw

References

- Emden R. Gansner and Stephen C. North, *An open graph visualization system and its applications to software engineering*, SOFTWARE - PRACTICE AND EXPERIENCE **30** (2000), no. 11, 1203–1233.

- MA Langston, AD Perkins, DJ Beare, RW Gauldie, PJ Kershaw, JB Reid, K Winpenny, and AJ Kenny, *Combinatorial algorithms and high performance implementations for elucidating complex ecosystem relationships from north sea historical data*, Proc. International Council for the Exploration of the Sea Annual Science Conference, 2006.

- Michael A. Langston and Brynn H. Voy, *Scalable Computational Methods for Analysis of the Low Dose Radiation Transcriptome*, Tech. Report TN 37996, Department of Computer Science, University of Tennessee, Knoxville, 2006.

- Shu Qin Li, Steve Guthridge, Edouard d'Espaignet, and Barbara Paterson, *From infancy to young adulthood: health status in the northern territory, 2006*, DHCS, 2007.

- Kerry-Ann O'Grady, Debbie Margaret Taylor-Thomson, Anne Bernadette Chang, Paul J Torzillo, Peter S Morris, Grant A Mackenzie, Gavin R Wheaton, Paul A Bauert, Margaret P De Campo, John F De Campo, et al., *Rates of radiologically confirmed pneumonia as defined by the world health organization in northern territory indigenous children*, Medical Journal of Australia **192** (2010), no. 10, 592–595.

- Taiyun Wei and Viliam Simko, *corrplot: Visualization of a correlation matrix*, 2016, R package version 0.77.

- Yun Zhang, Faisal N. Abu-Khzam, Nicole E. Baldwin, Elissa J. Chesler, Michael A. Langston, and Nagiza F. Samatova, *Genome-Scale Computational Approaches to Memory-Intensive Applications in Systems Biology*, SC '05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing (Washington, DC, USA), IEEE Computer Society, 2005, p. 12.

THE
END