Novel applications of Capacitated Cluster-Editing with Vertex Split Operation

FAISAL ABU-KHZAM, **PETER SHAW,** ALEXIS.SHAW, JUDITH EGAN, HEIDI SMITH-VAUGHA, ROBIN MARSH, NASIR JALAN & ANNE CHANG

UTS:QSI CENTRE FOR QUANTUM SOFTWARE AND INFORMATION

CENTRE FOR QUANTUM COMPUTATION & COMMUNICATION TECHNOLOGY
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

menzies school of health research

Abstract

The capacitated form of the cluster-editing produces significant improvements in the size of clinical data that can be processed. Nevertheless, the presence of (hub vertices in the network places undesirable constraints on the max-delete parameter. By introducing a vertex-split operation we can further constrain the parameter, allowing it to handle potentially larger data sets. The Capacitated Cluster-Editing with Vertex Split is NP-Hard and FPT. Our current project is exploring applications of this novel FPT problem in the search algorithm enables the analysis of Acute Respiratory Lung disease (ALRI) which is a major health issue in Australia's NT. Moreover, many other interesting applications exist and some of the will be briefly shared and discussed.

Overview

Motivation

Introduction

Cluster Editing and Cluster Editing with splits

Equivalence classes of edit sequences

Critical Clique Lemma

Kernelization

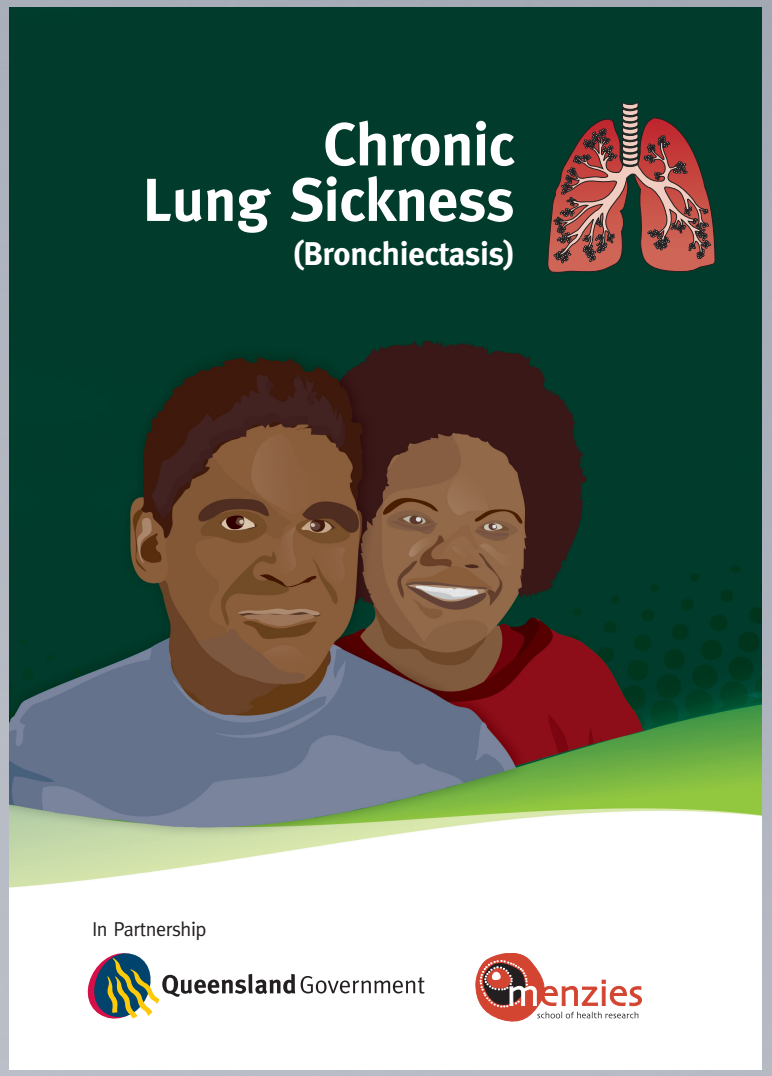
Conclusion

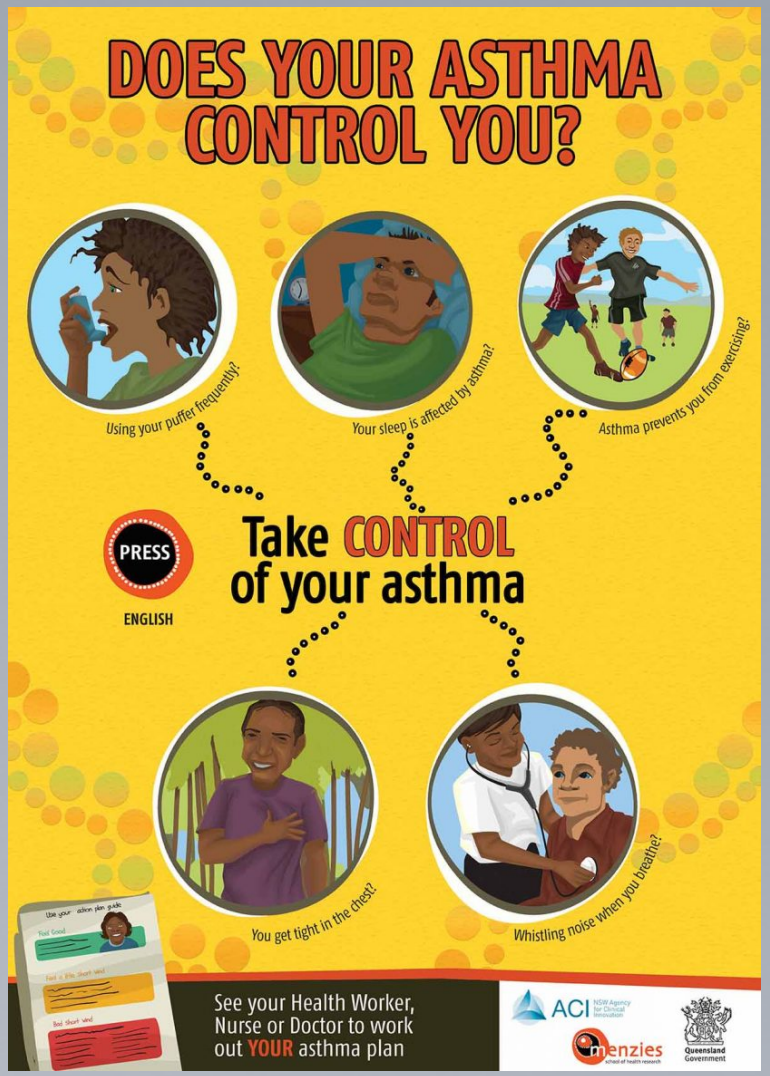
Motivation - Humane

- The motivation came from
  - Epidemiological
  - Clinical Research in NT Australia
- ARLI Pneumonia, Asma
  -
- Otitis Media (Ear Infections) Reflections

  30% of Children are deaf.

  99.7% aboriginal inmates are deaf.
- Currently vaccines and antibiotics are not sufficiently effective
- Multi-pathogen Diseases
- PCA and GLM limited to 1 : many interactions
- Clinical research requested Network model analysis





Various ALRI

Bronchiectasis often leads to other problems like ASTHMA

Motivation - Applied

Reflection on OM Study

- Better modelling techniques needed

Now Projects require larger Data Sets

- ARLI (other genetic) - 683+ vertices
- Diabetes Data - 20 vertices (target ok)
- Data Linkage Study (15000-50000 vertices)
- Social Networks (Twitter 100K-1M)

New Validation Techniques

- Novel Random Sampling and Boosting
- Disease ecology techniques like high-throughput sequence analysis.

Whats Needed

Applied

Need to be able to process bigger data sets

Need Better Models

- Cluster-Editing with Vertex Split (Inclusive)
- Capacitated Cluster-Editing (k,a,d)
- Capacitated Cluster-Editing (k,a,d,s)

More sophisticated Analytical Pipeline(s)

- Bayesian analysis
- Validation
  - Boosting/ Random Sampling
- Use of Machine Learning & Deep Learning in creating the Models

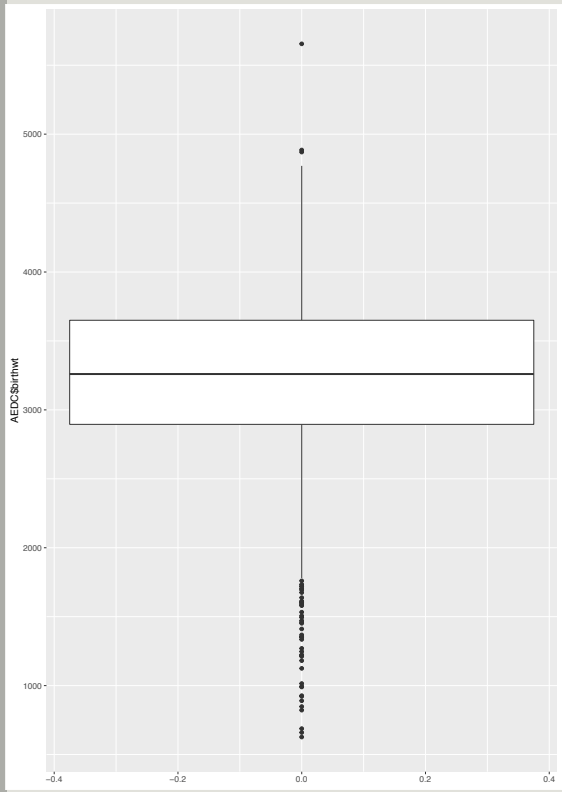
Type I & II Errors

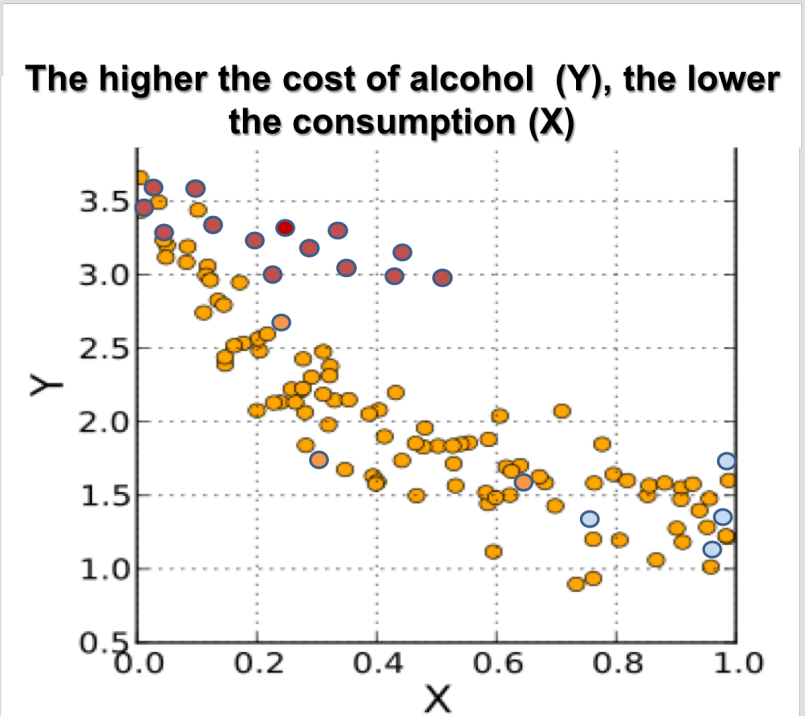
These are identified as added edges and deleted edges in the Cluster-Editing Search.

In the capacitated error we further restrict the local amount of Type I and II errors. This prevents unwanted trivial solutions such as isolating an important vertex.

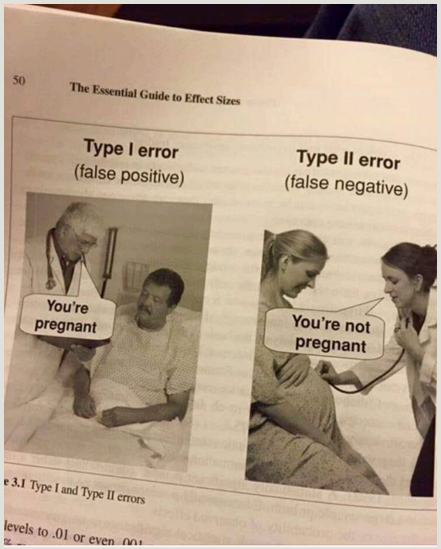
When two cliques share vertex(s) the d capacity needs to be set to high

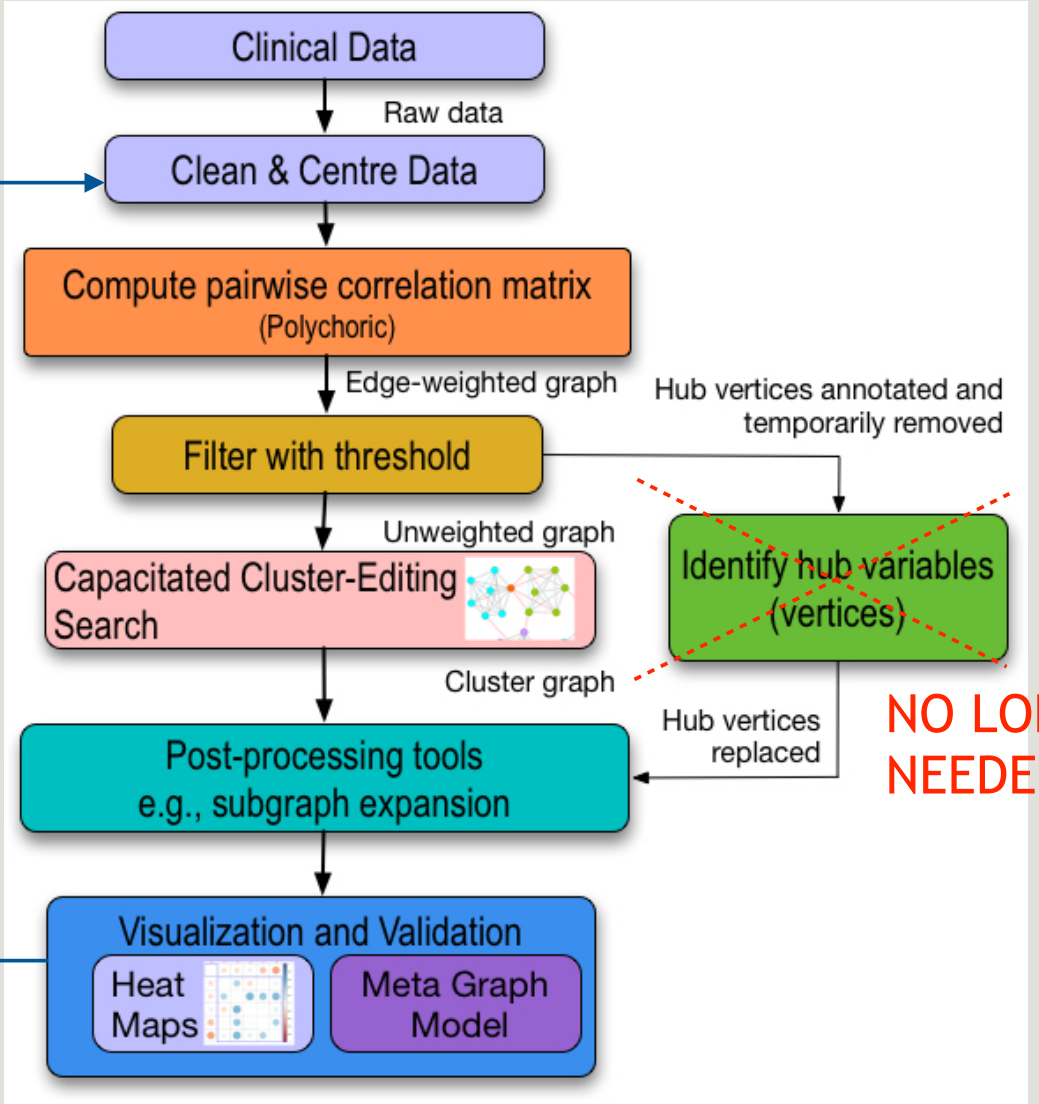
But this is more of a fudge than a real value for d.





Birth weights NT, This



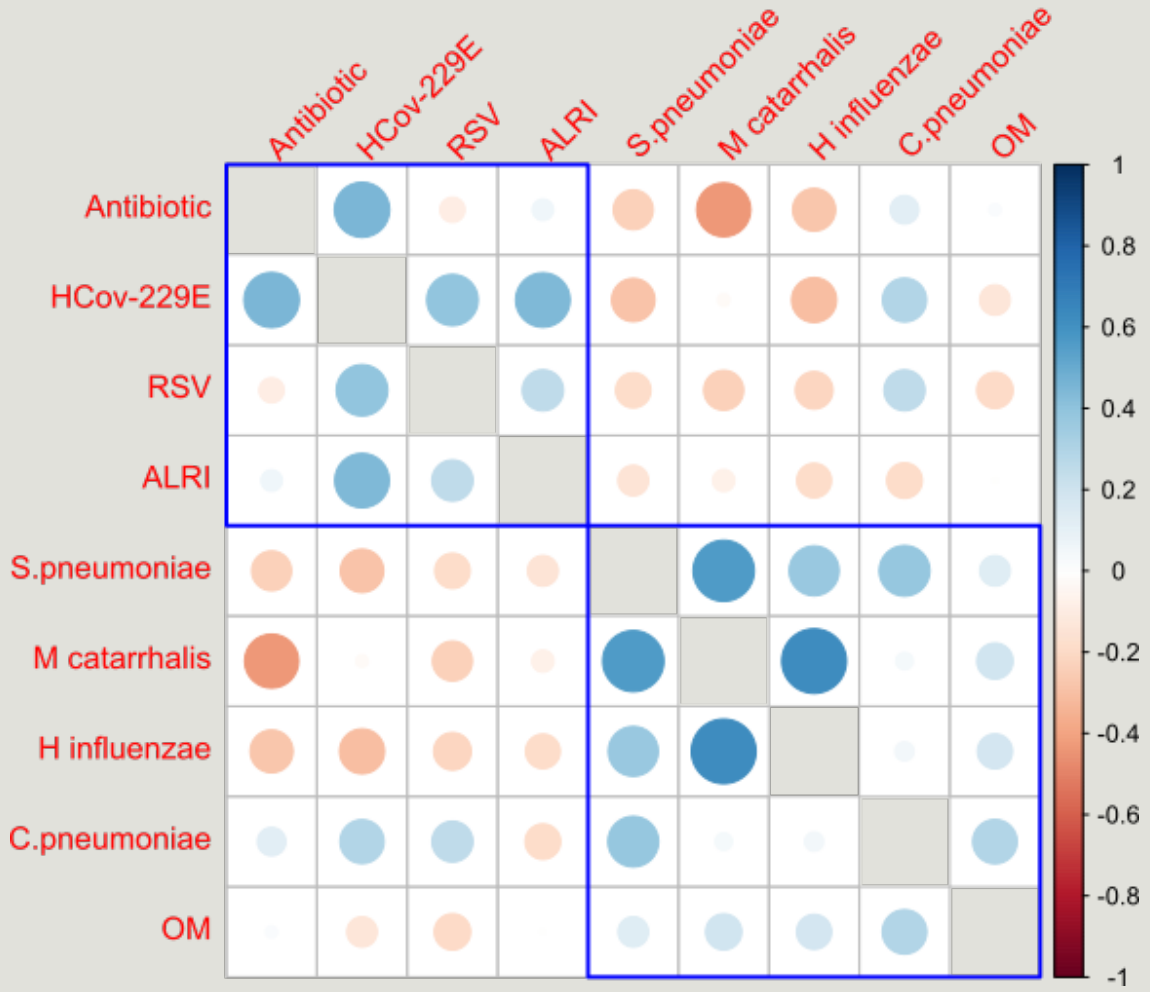


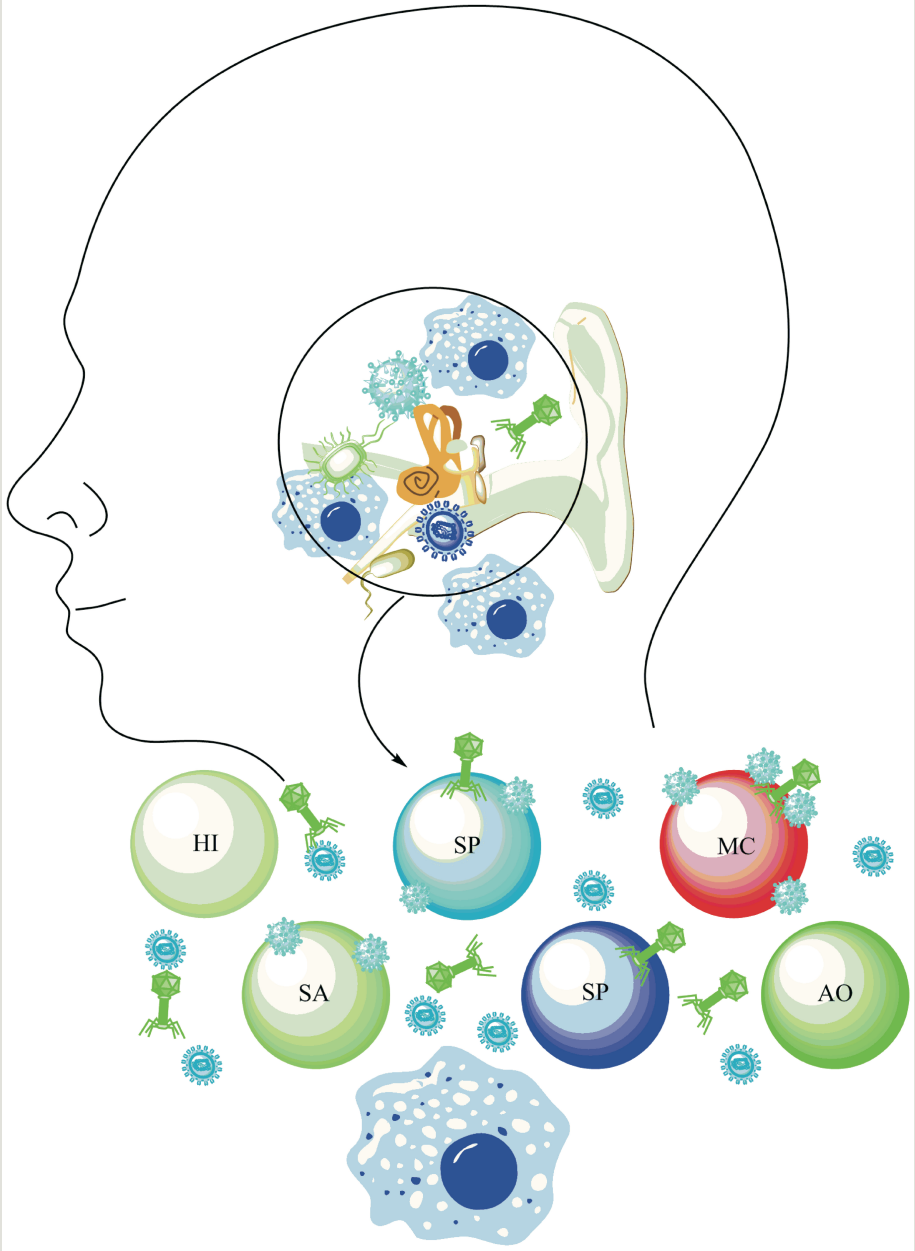
Clique Centric Pipeline

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

Searching for Para-clique Structures

A heatmap of Cluster 4 with secondary hysterical clusters (produced using R package corrplot [2])





Legend: Top shows a patient of Otitis media with a slew of micro-organisms that collectively cause a progressive form of chronic supparative otitis media (CSOM). Bottom part shows the same complex of disease ecology with a hypothetical illustration of viruses such as the Influenza virus, attacking the bacterial hosts *Hemophilus influenza* (HI) and *Moraxella catarrhalis* (MC) and *Streptococcus pneumonia* (SP), possibly causing them to become multi-drug resistant due to some form of genetic exchange or modification. Viruses in the ecosystem comprise respiratory syncytial virus, parainfleunza virus, influenza virus, entero virus, metapneumo virus, rhino virus and corona virus.

(Diagram by Nasir Jalal)

Abbreviations:

(HI) Hemophilus influenza;
(SP) Streptococcus pneumoniae,or Streptococcus pyogenes,
(MA) Moraxella catarrhalis;
(SA) Staphylococcus aureus;
(AO) Alloiococcus otidis.

Evidence for Hypothesis 1

Direct interactions with influenza promote bacterial adherence during respiratory infections.

Rowe HM[1], Meliopoulos VA[1], Iverson A[1], Bomme P[1,2], Schultz-Cherry S[1], Rosch JW[3].

Author information

Abstract

Epidemiological observations and animal models have long shown synergy between influenza virus infections and bacterial infections. Influenza virus infection leads to an increase in both the susceptibility to secondary bacterial infections and the severity of the bacterial infections, primarily pneumonias caused by Streptococcus pneumoniae or Staphylococcus aureus. We show that, in addition to the widely described immune modulation and tissue-remodelling mechanisms of bacterial-viral synergy, the virus interacts directly with the bacterial surface. Similar to the recent observation of direct interactions between enteric bacteria and enteric viruses, we observed a direct interaction between influenza virus on the surface of Gram-positive, S. pneumoniae and S. aureus, and Gram-negative, Moraxella catarrhalis and non-typeable Haemophilus influenzae, bacterial colonizers and pathogens in the respiratory tract. Pre-incubation of influenza virus with bacteria, followed by the removal of unbound virus, increased bacterial adherence to respiratory epithelial cells in culture. This result was recapitulated in vivo, with higher bacterial burdens in murine tissues when infected with pneumococci pre-incubated with influenza virus versus control bacteria without virus. These observations support an additional mechanism of bacteria-influenza virus synergy at the earliest steps of pathogenesis.

Capacitated Cluster-Editing with Vertex Splitting(CCECS)

Input: A graph $G = (V, E)$ and integers a,d,r and k;

Parameter: (k, a, d, r);

Question: Can G be transformed into a cluster graph with at most k edits (edges, additions or deletions and vertex splits) and furthermore:

- At most a edge additions are incident to any one vertex.
- At most d edge deletions are incident to any one vertex.
- r is the maximum number of edge splits ?

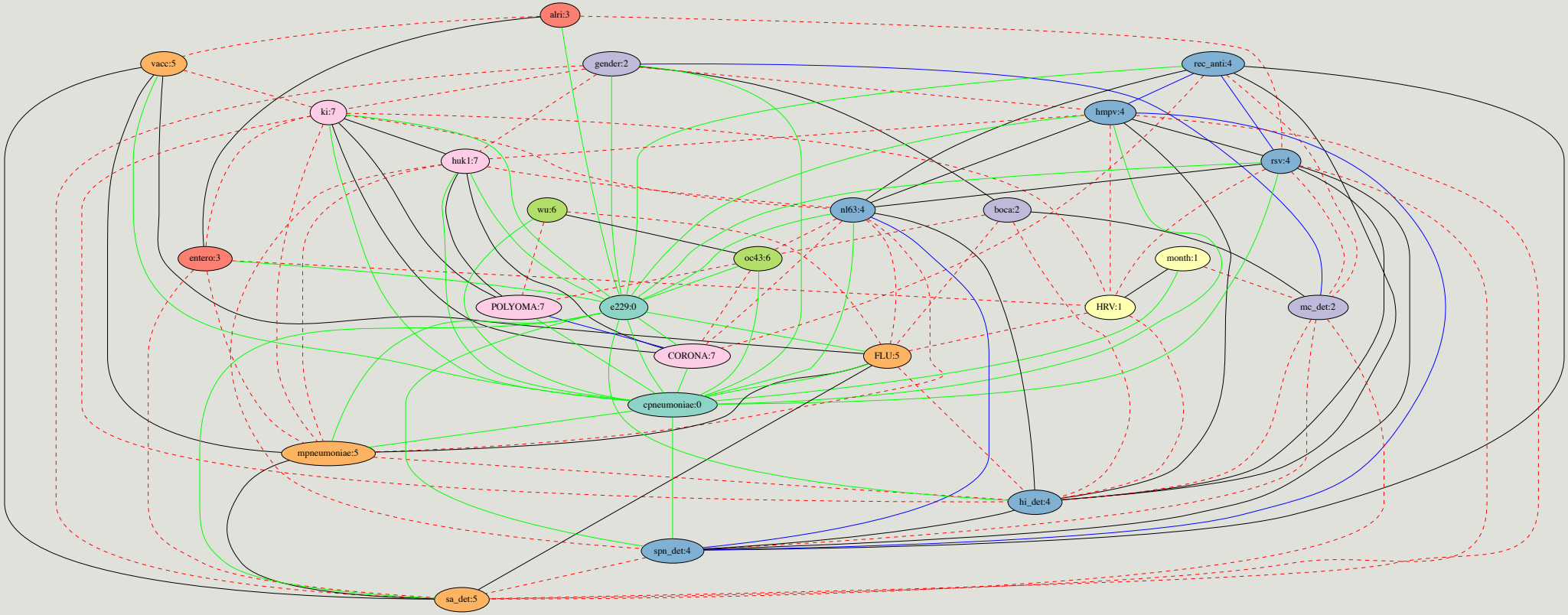
Definition 1

Cluster-Editing

Input:	A graph $G = (V, E)$ and integers k;
Parameter:	k;
Question:	Can G be transformed into a cluster graph with at most k edits (edges, additions or deletions)?

This problem is well studied in the literature, for example by Ben-Dor et. al. [1] and Dehne et. al. [6]

Cluster-Edit solution A 77 edits - **but two solutions existed**



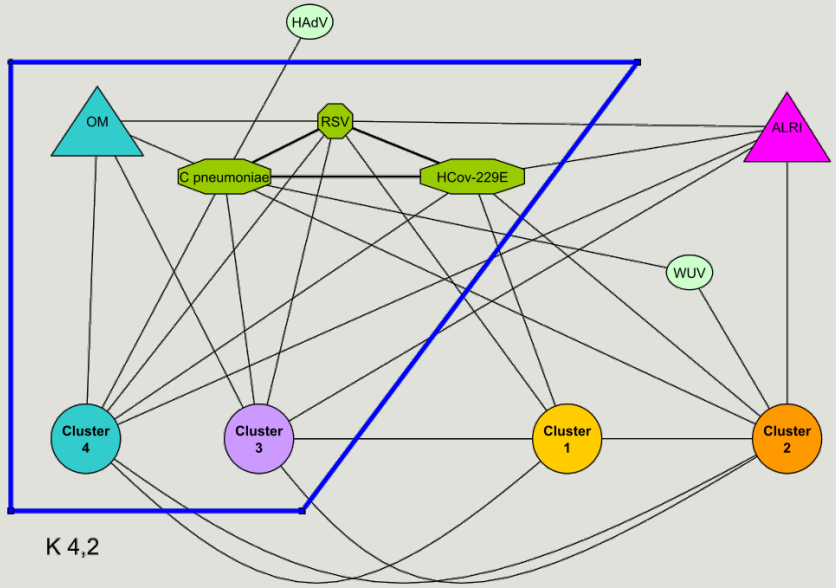
Clinical

Key feature are vertices shared by two or more cliques

But high degree (aka hub) vertices raise two issues

1. Which clique do they belong to. i.e. after the edges are removed
2. The number of edges removed for each hub vertex can be $O(n)$

As $T(n,k) = O^*\left(1.62^k\right)$ handling these high degree vertices is expensive



Definition 2
Cluster Editing with Vertex Splitting (CEVS)

Instance A graph $G = (V, E)$ and an integer k

Parameter k

Question Can G be transformed into a graph G' that is a disjoint union of cliques by at most k of the following operations:

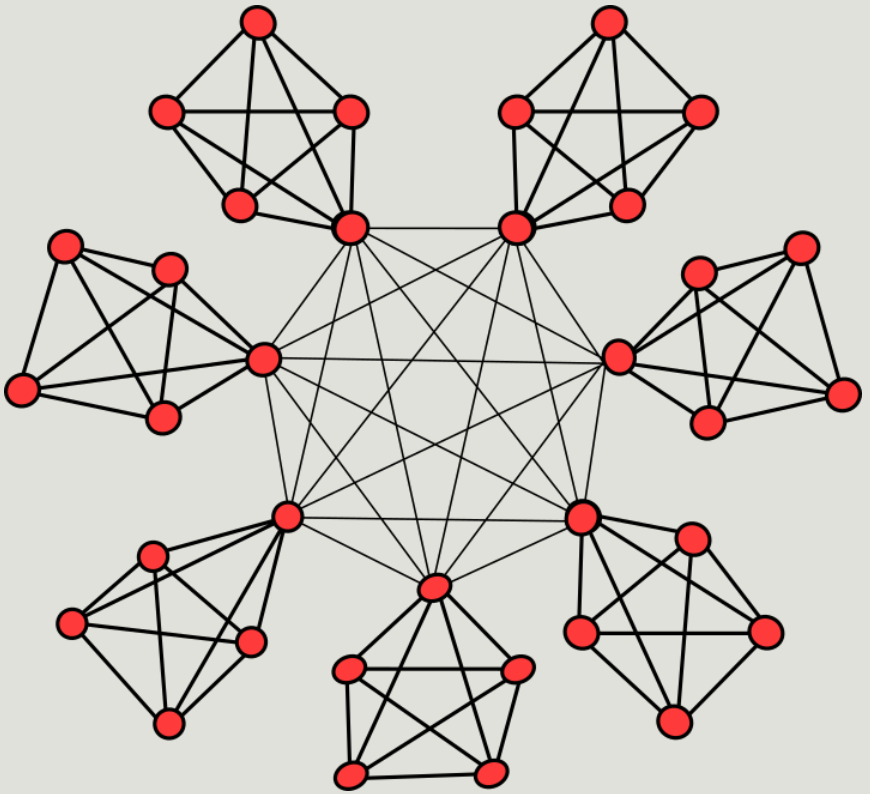
1. The identity operation (does nothing)
2. The addition of an edge to E
3. The deletion of an edge from E
4. An inclusive vertex splitting

Why add splits

Cluster editing works well if a graph has non-overlapping clusters, however if cliques overlap in even a few vertices then then clusters can be lost.

On the graph right, the cluster editing problem would remove the K_7 in the middle with an edit cost of 20.

Adding splits allows efficient solutions to cases where there are overlapping clusters, and maintains structure in the solution.



Construction from paper by Damaschke [2]

Definition 3

Capacitated Cluster-Editing with Vertex Splitting(CCECS)

Input: A graph $G = (V, E)$ and integers a,d,r and k;
Parameter: (k, a, d, r);
Question: Can G be transformed into a cluster graph with
at most k edits (edges, additions or deletions and vertex splits)
and furthermore:

- At most a edge additions are incident to any one vertex.
- At most d edge deletions are incident to any one vertex.
- r is the maximum number of edge splits ?

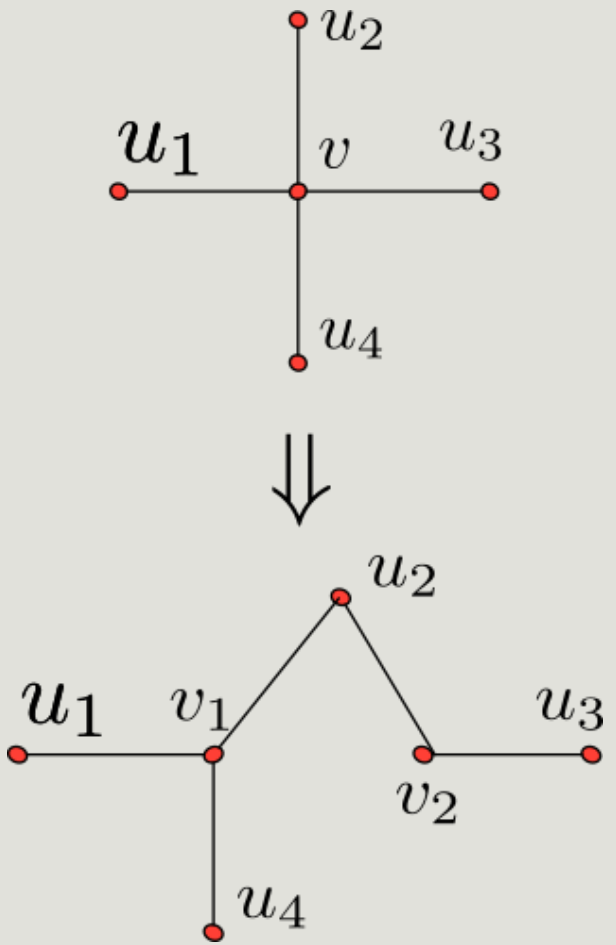
Lemma 1. *Capacitated Cluster Editing with Vertex Splitting (CCEVS) is in* NP*–Hard.*

Proof. CCEVS is trivially NP-Hard by special case reduction. As both the Cluster Deletion and Cluster Vertex-Deletion problems [1] are both NP-hard. Just set $a = 0, d = 0$ or $s = 0$ □

Inclusive vertex splitting

The inverse of vertex merger, more formally:

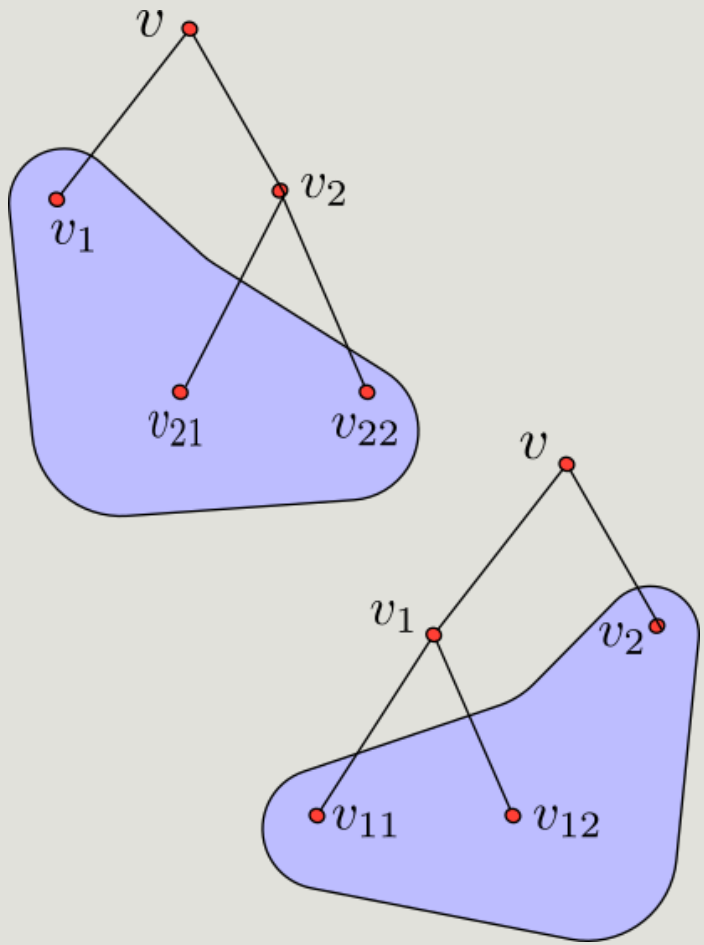
- For come $v \in V$ an *inclusive vertex splitting* is defined as a partitioning of the vertices in $N(v)$ into two disjoint sets $U_i,\ U_2$ such that $U_1 \cup U_2 = N(v)$ and then removing v from the graph and adding two new vertices v_1 and v_2 with $N(v_1) = U_1$ and $N(v_2) = U_2$



Vertex split relations

For every edit sequence $S = e\_1 \ldots e\_k$ the vertices of G_S can be seen as being related to the vertices of G by a binary tree known as the split-tree for v where if a vertex v is split by some e_i then it is the parent to the two new vertices u and w each vertex of G_S is a leaf node of some tree and each vertex of G is a root.

We define the *Vertex split relation* $R: V \to 2^{V_S}$ such that $R_S(v)$ is the set of leaf nodes of v's split-tree



Whats Needed

Theoretical

A. Smaller Kernels

B. Faster (Randomised) Algorithms

C. Multi-Parameterization

- In practice multi-parameterisation means 1000X speedup and much smaller kernels

D. New Data Structures

- Hybrid Data Structures [7,8] expanded for cluster-editing
- New w.i.p. [10]

$$R^D \qquad (u,v) \in E \text{ if } d(u,v) \leq r, \quad (u,v) \notin E \text{ if } d(u,v) > R$$

E. HPC Algorithms/Implementations

- GPU
- Clusters
- BOTH

F. New Applications

- SAT, LOGIC
- Bayesian Analysis

Here Lemma 10 from [7b] doesn't necessarily apply for the multi-parameter version of the problem.

Lemma 10. *If there is a solution to* CEVS *on* (G, k) *then there are at most* $4k$ *non-isolated vertices in* $CC(G)$. *Moreover, there are at most* $3k + 1$ *vertices in any connected component of* $CC(G)$ *and there are at most* k *connected components in* $CC(G)$ *which are non-isolated vertices.*

Nerveless, by the main results of Lemma 6 and 7 do

Lemma 6. *For any collection of edit-sequences in add-delete-split form which are equivalent to some edit-sequence* S, *the graph* G_{R_S} *immediately preceding the vertex splitting is the same for all members of the class in that form. Further if the split relation for this equivalence class is* R *and the graph* $G' = (V'E')$ *resulting from* S *are known, then*

$$E' = \{u, v \in V' : \exists u' \in R(u)\ \exists v' \in R(v) \text{ such that } u'v' \in E\}\ .$$

Lemma 7. *For any graph* $G = (V, E)$ *there is a computable bijection between pairs* $(G' = (V', E'), f : V \to 2^{V'})$ *of resultant graphs and split-relations and equivalence classes of edit-sequences. Further there is an algorithm to compute a min-edit-sequence from the resultant graph/split-relation pair for this class in* $\mathcal{O}((|V'| - |V|)\Delta(G) + |V| + |E| + |V'| + |E'|)$ *time.*

Nevertheless, Theorem [7b] 2 still apply as the conversion from can only reduce each of the parameters individually. And so there there is a canonical representation by covering that can be applied.

Theorem 2. *There is a bijection between coverings and equivalence classes of edit-sequences. Furthermore, we can compute a canonical (minimum like) sequence from the equivalence classes in* $\mathcal{O}((|V'|-|V|)\Delta(G)+|V|+|E|+|V'|+|E'|)$ *time.*

From this we can obtain a FPT kernel and so CCEVS is FPT but this kernel could be improved.

Here Lemma 10 doesn't necessarily apply for the multi-parameter version of the problem.

Lemma 10. *If there is a solution to* CEVS *on* (G,k) *then there are at most* $4k$ *non-isolated vertices in* $CC(G)$*. Moreover, there are at most* $3k+1$ *vertices in any connected component of* $CC(G)$ *and there are at most* k *connected components in* $CC(G)$ *which are non-isolated vertices.*

References

[1] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, “Tissue classification with gene expression profiles,” *J Comput Biol*, vol. 7, no. 3-4, pp. 559–583, 2000. 51, 222

[2] P. Damaschke, “On the fixed-parameter enumerability of cluster editing,” in *Proceedings, International Workshop on Graph-Theoretic Concepts in Computer Science*, ser. Lecture Notes in Computer Science, D. Kratsch, Ed., vol. 3787. Springer, 2005, pp. 283–294

[3] Lin, G.H., Kearney, P.E., Jiang, T.: *Phylogenetic k-root and Steiner k-root*. In: Algorithms and Computation, pp. 539–551. Springer (2000)

[4] Guo, J.: *A more effective linear kernelization for cluster editing*. Theoretical Computer Science 410(8-10), 718 – 726 (2009)

[5] M Fellows, M Langston, F Rosamond, P Shaw: *Efficient parameterized preprocessing for cluster editing* International Symposium on Fundamentals of Computation Theory, 312-321

[6] F Dehne, MA Langston, X Luo, S Pitre, P Shaw, Y Zhang: *The cluster editing problem: Implementations and experiments* International Workshop on Parameterized and Exact Computation, 13-24

[7] Abu-Khzam, Faisal N., et al. "Cluster editing with vertex splitting." International Symposium on Combinatorial Optimization. Springer, Cham, 2018.

[7b] Abu-Khzam, Faisal N., et al. "Cluster editing with vertex splitting." Discrete and Applied Maths 2019 (submitted).

[8] Abu-Khzam, Faisal N., et al. "A hybrid graph representation for recursive backtracking algorithms." International Workshop on Frontiers in Algorithmics. Springer, Berlin, Heidelberg, 2010.

[9] Abu-Khzam, Faisal N., et al. "Accelerating vertex cover optimization on a GPU architecture." 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). IEEE, 2018.

[10] Abu-Khzam, Faisal N., and Rana H. Mouawi. "Concise Fuzzy Representation of Big Graphs: a Dimensionality Reduction Approach." arXiv preprint arXiv:1803.03114 (2018).

END