Integrated simultaneous analysis of different biomedical data types with exact weighted bi-cluster editing

Peng Sun, Jan Baumbach, Jiong Guo

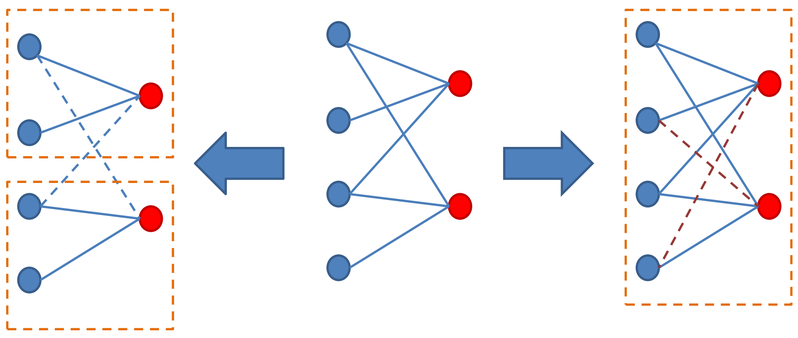
Computational Systems Biology Group
Max Planck Institute for Informatics,
Cluster of Excellence, University of Saarland
Address correspondence: psun@mpi-inf.mpg.de

Abstract

- The explosion of biological data has largely influenced the focus of today's biology research. Integrating and analysing large quantity of data to provide meaningful insights has become the main challenge to biologists and bioinformaticians. One major problem is the combined data analysis of data from different types, such as phenotypes and genotypes. Here we contribute with an exact algorithm that is based on fixed-parameter tractability.

Bi-Cluster Editing

- Given a graph $G = (V, E)$, can we convert this graph into cliques with at most k edge modifications (or with modification penalty at most k)?



Exact Algorithms

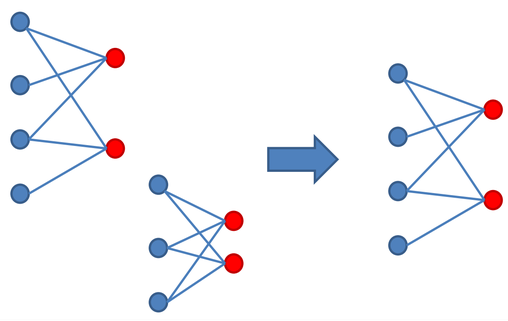
Fix Parameter Approach

- NP-hard problems are computable in a time that is **polynomial of input size** and **exponential or worse in a parameter *k***.

Kernelization

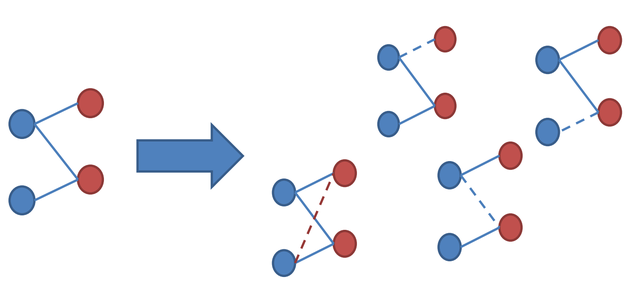
- A kernelization is an **efficient mapping of the input instances into equivalent instances with a guaranteed upper bound** on the size.

- $(x, k) \mapsto (x', k')$
- $O(x') = f(k)$
- $O(k') = g(k)$



Branching Strategy

- Branching strategy is the approach of the depth-first searching tree to solve the problem.

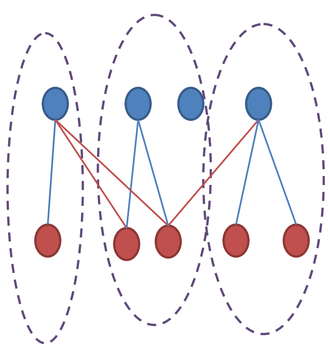


We can design faster kernelization and branching strategy.

Results (I)

Artificial Graphs

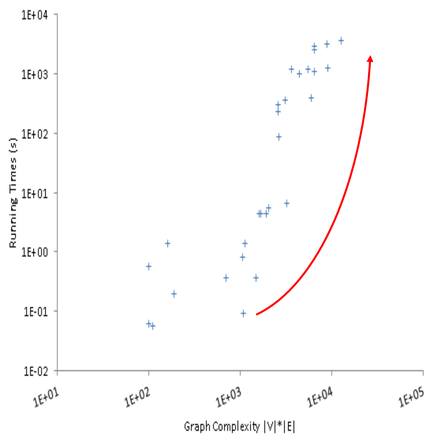
- We generate random graphs with given vertices and random assigned edge weights.
- Two Gaussian distributions are used to generate the edge weights.



Results (I)

The exponential explosion

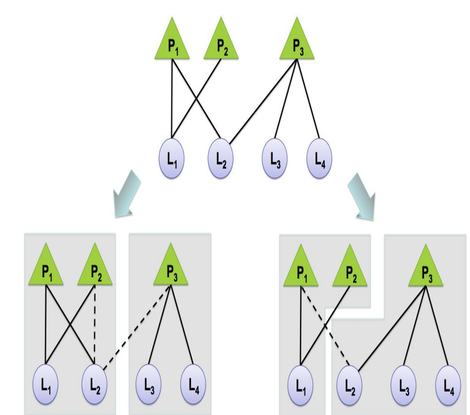
- The figure on the right visualize the running times against the graph component complexity.



Results (II) : on GWAS data

(a) Proposal

- **Vertices**: (1) loci/genotypes
  Edges: significant associations
  (2) phenotypes
- To identify **groups of variations** responsible for **groups of diseases**.



(b) DATA SOURCES

	Literature Search	NHGRI Dataset
Number of Associations	54,776	4,325
Number of SNP loci	52,644	3,949
Number of Phenotypes	87	414

(c) Putative Associations

- We identified 86 putative associations.

Traits/Disease	No. of Newly Found Associations
Conduct disorder (case status) *	11
Ischemic stroke	11
Atrial fibrillation/atrial flutter*	10
Permanent tooth development*	10
Conduct disorder (symptom count) *	9
Primary tooth development (time to first tooth eruption) *	8
Cleft lip*	7
Primary tooth development (number of teeth) *	5
Alcoholism (alcohol dependence factor score)*	4
Plasma coagulation factors*	3
Vitamin D insufficiency*	3
Vitamin D levels*	2
Atrial fibrillation*	1
Nonsyndromic cleft lip with or without cleft palate*	1
Plasma levels of Protein C*	1
Total	86

References:
S. Rahmann *et al.* Exact and heuristic algorithms for weighted cluster editing.
N. Amit. The Bicluster Graph Editing Problem. M. Sc thesis. (2004)
F. Protti *et al.* Applying Modular Decomposition to Parameterized Bicluster Editing. (2006) IWPEC
J. Guo *et al.* Improved Algorithms for Bicluster Editing. (2008) PROC. 5TH TAMC LNCS
R. G. Downey and M. R. Fellows. Parameterized Complexity. Springer. (1999)

ACKNOWLEDGEMENT:
All authors wish to thank the Cluster of Excellence for Multimodal Computing and Interaction of the German Research Foundation for financial support.