On Bipartite Graph Decomposition in the Presence of Noise, with Applications to Biological Data Clustering

Conference Paper · January 2012

CITATIONS
0

READS
202

5 authors, including:

Erich J Baker
Baylor University
97 PUBLICATIONS **363** CITATIONS

SEE PROFILE

Elissa J Chesler
The Jackson Laboratory
239 PUBLICATIONS **7,490** CITATIONS

SEE PROFILE

Michael A. Langston
University of Tennessee
313 PUBLICATIONS **4,818** CITATIONS

SEE PROFILE

On Bipartite Graph Decomposition in the Presence of Noise, with Applications to Biological Data Clustering

Charles A. Phillips[1], Jeremy J. Jay[2], Erich J. Baker[3],
Elissa J. Chesler[2] and Michael A. Langston[1]

[1] Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville, TN 37996, USA
{cphillip, langston}@eecs.utk.edu
[2] The Jackson Laboratory, Bar Harbor, ME 04609, USA
{jeremy.jay, elissa.chesler}@jax.org
[3] School of Engineering and Computer Science,
Baylor University, Waco, TX 76798, USA
erich_baker@baylor.edu

Keywords: paraclique, biclique, parabiclique, biclustering, noise-resilient algorithms

1 Overview

We present a novel algorithm for extracting dense, disjoint subgraphs from undirected bipartite graphs. Our procedure successively removes such subgraphs, known as parabicliques, by iteratively isolating a maximum biclique and then expanding it in the presence of missing edges. Hence it relies on our previous work on efficiently finding solutions to the $\mathcal{NP}$-complete maximum biclique problem. It is also resilient to noise in the form of outliers, poorly correlated raw data and so forth. We have implemented the algorithm and tested it on heterogeneous biological graphs that represent, among other things, associations between genes and diseases, phenotypes, and even microbes. This approach to biological data analysis can be employed as a tool for discovering, confirming and hypothesizing the many roles of genes, gene products and a wide variety of other biological network agents.

2 Background

Bipartite graphs provide a natural way to model associations between pairs of heterogeneous object classes. Maximally connected subgraphs in bipartite graphs, called bicliques, have proved useful in a huge array of application domains, from computational biology [1, 2] to wireless networks [3]. Hosts of algorithms address the related problem of biclustering, or co-clustering, in which the rows and columns of a matrix are clustered simultaneously [4, 5]. A two-dimensional matrix can be interpreted as a weighted bipartite graph, and so finding bicliques in undirected graphs can be viewed as a special case of biclustering in which matrix entries are binary.

A *paraclique* is a subgraph consisting of a maximum clique along with any additional vertices that meet some pre-defined connectivity criteria. Algorithms for paraclique construction have been described and applied in several previous venues. See,

for example, [6]. Paraclique is rather ideally suited to decomposing high dimensional data and ameliorating the effects of noise. Here we extend the paraclique notion to the problem of effective bipartite graph clustering. Informally, a *parabiclique* is a maximum biclique augmented with additional vertices that preserve high but not perfect density. As with paraclique, the main motivations are to decompose highly overlapping edge sets and provide effective data clustering in the presence of noise.

3 Edge Maximum and Vertex Maximum Bicliques

A *maximal* biclique B is one to which no vertex can be added to form a larger biclique. That is, B is not properly contained in any other biclique. A *maximum* biclique is the largest biclique in a graph. For both maximal and maximum bicliques, we must distinguish between two variants: vertex-maximal (or maximum) and edge-maximal (or maximum). In the former, the size of a biclique is its number of vertices; in the latter, the size is the number of edges. Figure 1 illustrates the difference. More telling from an algorithmic standpoint, the vertex-maximum biclique in a graph can be found in polynomial time, but the problem of finding the edge-maximum biclique is $\mathcal{NP}$-complete [7].

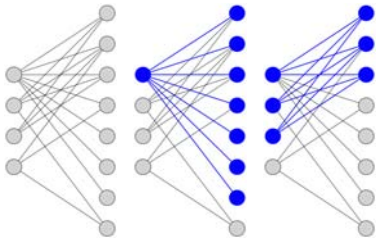

Fig. 1. The bipartite graph in (a) is shown in (b) with its vertex-maximum biclique highlighted in blue and in (c) with its edge-maximum biclique highlighted in blue.

In practice, edge-maximal biclique is generally the more interesting of the two problem variants. It is probably not surprising that it is the $\mathcal{NP}$-complete problem that is the more useful. Edge-maximal bicliques tend to contain multiple vertices from each class, and thus have a higher ratio of edges to vertices and more relationships per vertex. In the domain of biological data analysis, this translates to more relevant molecular response networks and other sorts of putative functional modules. Fortunately, an asymptotically efficient algorithm for enumerating all edge-maximal bicliques, and one that we will use in our implementation of parabiclique, can be found in [8].

4 Algorithm

We begin by finding a maximum biclique, B. Every vertex not contained in B is then evaluated. If a vertex has sufficient connectivity to B, it is added to B. Otherwise it is

discarded. Connectivity to B is generally determined using two parameters, g and h. For historical reasons, these are termed *glom factors*, one for each vertex class. We interpret glom factors in one of two ways: they either denote the number of edges allowed to be missing, or the proportion of edges that must be present. (In the latter case, parameter are called *proportional* glom factors. Figure 2 displays an example parabiclique. In order to handle cases for which the maximum biclique contains only a few representatives from one class, we sometimes include two additional parameters, w and x, which specify the minimum number of vertices a maximum biclique must contain from each class. (If the biclique contains fewer vertices, then vertices from the other class are not considered for inclusion.) Adjusting parameters g, h, w, and x allows the algorithm to be fine-tuned to bipartite graphs from different application domains. As just one example, testing gene-geneset graphs from Gene Weaver [9] revealed that a proportional glom factor of 0.25 was not unreasonably low, because two genesets having that proportion of common genes are significant.

Input: A bipartite graph G with partitions U and V, glom factors g, h, and parameters w, x
Output: A parabiclique, P
$P = B$ = Maximum biclique in G, with partitions $W \subseteq U$ and $X \subseteq V$;
if $|W| \geq w$ **then**
  foreach $v \in V \setminus X$ **do**
    if v *is connected to at least* $g|W|$ *vertices in* W **then**
      $P = P \cup \{v\}$;
    end
  end
end
if $|X| \geq x$ **then**
  foreach $v \in U \setminus W$ **do**
    if v *is connected to at least* $g|X|$ *vertices in* X **then**
      $P = P \cup \{v\}$;
    end
  end
end
return P;

Algorithm 1: Extracting a single parabiclique from a bipartite graph G. Disjoint parabicliques are iteratively extracted by setting $G = G \setminus P$ with each successive call to the algorithm. Iteration continues until some stopping condition is reached, typically when a predetermined number of parabicliques is extracted or when $|B|$ falls below some value. Shown is the proportional glom factor variant. For a simple glom factor version, replace $g|X|$ and $g|W|$ with $|X| - g$ and $|W| - g$ respectively.

For efficiency and scalability, we employ the powerful biclique enumeration algorithm as described in [8]. Once a starting maximum biclique has been identified, vertex connectivity computations require at most quadratic time.

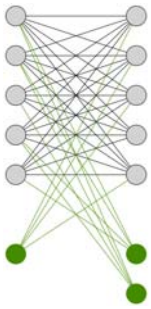

Fig. 2. A parabiclique. Vertices of the maximum biclique are in grey. Green vertices are missing just one edge to vertices in the opposing class, thus are included in the parabiclique when $g \geq 1$.

Preliminary applications have been revealing. In gene-geneset graphs from Gene Weaver, for example, we have been able to identify novel and previously unknown associations between disparate experiments using this algorithm. CTW submission limitations prohibit a thorough discussion of such compelling results. If this paper is accepted, we will address them in detail during the workshop talk.

References

1. Pati, P., Vasquez-Robinet, C., Heath, L., Grene, R., Murali, T.M.: XcisClique: Analysis of Regulatory Bicliques. BMC Bioinformatics. 7:218 (2006)
2. Schweiger, R., Linial, M., Linial, N.: Generative Probabilistic Models for Protein-Protein Interaction Networks-The Biclique Perspective. In: Bioinformatics 27 (13): i142-i148 (2011)
3. Zhong-Ji, F., Ming-Xue, L.; Xiao-Xin, H.; Xiao-Hui, H.; Xin, Z.: Efficient Algorithm for Extreme Maximal Biclique Mining in Cognitive Frequency Decision Making. In: IEEE 3rd International Conference on Communication Software and Networks (2011)
4. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms For Biological Data Analysis: A Survey. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 1, no. 1, pp. 24-45 (2004)
5. Tanay, A., Sharan, R., Shamir, R.: Biclustering Algorithms: A Survey. In: Chapman, A.S. (ed.), Handbook of Computational Biology. Hall/CRC Computer and Information Science Series. (2005)
6. Chesler, E.J., Langston, M.A.: Combinatorial Genetic Regulatory Network Analysis Tools for High Throughput Transcriptomic Data. In: RECOMB Satellite Workshop on Systems Biology and Regulatory Genomics (2005)
7. Peeters, R.: The maximum edge biclique problem is $\mathcal{NP}$-complete. In: Discrete Applied Mathematics, 131(3):651.654 (2003)
8. Zhang, Y., Chesler, E.J., Langston, M.A.: On Finding Bicliques in Bipartite Graphs: A Novel Algorithm with Application to the Integration of Diverse Biological Data Types. In: Proceedings, Hawaii International Conference on System Sciences, Big Island, Hawaii (2008)
9. Baker, E.J., Jay, J.J., Bubier, J.A., Langston, M.A., Chesler, E.J.: GeneWeaver: A Web-based System for Integrative Functional Genomics. In: Nucl. Acids Res. 40(D1): D1067-D1076 (2012)