Clustering without any subjective similarity measure - Regularity Clusterings

*Shai Ben-David*
University of Waterloo
*Joint work with Sharon Wulff*

*Non-Metric Learning Workshop*
Haifa, 2010
A different clustering task

- Consider the task of clustering university web pages based on the graph of links between these pages.
- Can clusters of “functionally similar” pages be detected from just this link structure?
Differences from common clustering tasks

- There is no input similarity (or distance measure) between the domain elements. All the information in the input comes as objective, observed, binary relations among the objects.

- These relations are not similarity links. For example, the cluster of professors’ pages may have very few internal links.
The clustering goal (intuitively)

What we are looking for are clusters whose members share similar link patterns with respect to members of the other clusters.

Example from systems biology:

“Functional classification of drugs by properties of their pairwise interactions”
Epistasis Networks

[Kishony et al, Nature genetics 2007]

- Networks that describe how perturbations (mutations or drugs) in a given biological system are combined to aggravate or alleviate the phenotypic consequences of each other.

- “Epistasis networks posses a special property, which we term “monochromaticity”, whereby functional gene modules interact with each other with purely aggravating or purely alleviating epistatic links.”

- In drug networks, the same conceptual method allow classification of drugs by their mechanism of actions based only on the properties of their mutual interactions.
Before we can discuss algorithms and computational complexity, we need to formally define what are we trying to achieve.

**Input:** A symmetric matrix, $M \in \{0,1\}^{m \times m}$, and a number, $k$.

**Output:** A partition, $C_1, \ldots C_k$ of $\{1,\ldots m\}$ minimizing the following objective function (s):
For a matrix $B$, let $\text{Maj}(B)$ denote the most common entry value in $B$.

The minority cost of a partition $C_1, \ldots, C_k$ of $\{1, \ldots, m\}$ is

$$\sum_{i \leq k, j \leq l} \left| \{(a, b) \in C_i \times C_j : M(a, b) \neq \text{Maj}(C_i \times C_j)\} \right| \frac{|C_i| \times |C_j|}{|C_i| \times |C_j|}$$
Interpretations

- The Minority clustering can be thought of as a compression of the matrix, replacing $mxm$ entries by $kxk$ block entries.

- The minority cost equals the error of that compression in predicting the entries.
The variance objective function

- For a vector $v \in \{0,1\}^d$
  
  Let $p(v)$ denote its average.

- For a matrix $M$ as above and a partitioning $C=(C_1, \ldots, C_k)$ of $\{1,\ldots,m\}$, define

$$O_{\text{var}}(C) = \sum_i \sum_j \text{Variance} \{p(v) : v \text{ is a row in } C_i \times C_j\}$$
Relationship to Szemeredi’s Regularity Lemma

A fundamental result in graph theory, Szemeredi’s regularity lemma, can be viewed as stating that,
For every $\varepsilon$, for every large enough $k$ and $m$ (a huge function of $\varepsilon$),

*every input matrix has an $\varepsilon$–regular $K$-clustering (even with equal size clusters).*
Correlation Clustering

**Input:** Complete graph with -1/1 labels on its edges \((V, L)\) and a number \(k\).

**Output:** Partition of the set of nodes \(V\) into \(k\) subsets (clusters).

**Cost:** The sum of within clusters -1’s plus the sum of between clusters 1’s.
Connection to Regularity Clustering

Note that *correlation clustering* can be thought of a special case of *Minority Cost* biclustering, where:

- *the input matrix is symmetric*,
- *the output clustering of rows equals that of columns*, and
- *the block labels are 1 on the diagonal and -1 on all other blocks.*
Extensions to Co-Clustering

- One can apply the same objective functions to bi-clustering (or co-clustering) problems.

- Co-Clustering:
  
  The input is an \( nxm \) matrix and two parameters, \( k,l \),

  The output: hPartitions, \( C_1, \ldots C_k \) of \( \{1,\ldots m\} \)
  
  \( R_1, \ldots R_l \) of \( \{1,\ldots n\} \)
We wish to cluster movies, and cluster viewers, based (only) on the ranking table data.
Bi-clustering of Gene expression micro-arrays

Figure 1.1: A sample of bi-clustering in computational biology. The left figure shows a gene expression data matrix, and the right figure shows the bi-clustering result on it.
The Computational Complexity of Regularity Clusterings

1. Optimizing any of the above objectives is NP-hard.

2. There exist PTAS algorithms for each of these objective functions.
1. Randomly chose a sample $S$ of points of size depending only of the desired approximation quality (independent of $m$).

2. For each $k$-clustering of $S$, find the clustering it induces on the full domain $\{1, \ldots m\}$. The notion of induced clustering depends on the objective function at hand.

3. Calculate the objective cost of each of these induced clustering and output the $\text{argmin}$. 
Conclusions:

- Regularity clustering works on input without a metric (without any prior similarity measure).
- We propose several precise objective functions for “regularity clustering” tasks.
- We provide computational complexity analysis of these tasks, as well as approximation algorithms with performance guarantees.
- Although these algorithms are formally poly-time, they are too expensive to run in practice.
Future research direction

- Develop algorithms that can be run in practice …
- *Use such clusterings to deduce a metric (or similarity measure) over the domain objects.*