Szemerédi's Regularity Lemma and Pairwise Clustering

Anna Sperotto
University of Twente, The Netherlands

Marcello Pelillo
University of Venice, Italy

Talk's Outline

- Szemerédi's regularity lemma
- Finding regular partitions in polynomial time
- The regularity lemma and pairwise clustering
- Experimental results

Talk's Outline

- Szemerédi's regularity lemma
- Finding regular partitions in polynomial time
- The regularity lemma and pairwise clustering
- Experimental results

Edge Density

Let $G = (V, E)$ be an undirected graph with no self-loops, and let $X, Y \subseteq V$ be two disjoint subsets of vertices of G.

We define the edge density of the pair (X, Y) as:

$$d(X, Y) = \frac{e(X, Y)}{|X||Y|} \tag{1}$$

where $e(X, Y)$ denotes the number of edges of G with an endpoint in X and an endpoint in Y, and $|\cdot|$ denotes the cardinality of a set.

Note that edge densities are real numbers between 0 and 1.

Regularity

Given a positive constant $\varepsilon > 0$, we say that the pair (A, B) of disjoint vertex sets $A, B \subseteq V$ is ε-regular if for every $X \subseteq A$ and $Y \subseteq B$ satisfying

$$|X| > \varepsilon|A| \quad \text{and} \quad |Y| > \varepsilon|B| \tag{2}$$

we have

$$|d(X, Y) - d(A, B)| < \varepsilon \ . \tag{3}$$

Thus, in an ε-regular pair the edges are distributed fairly uniformly.

Equitable Partitions

A partition of V into pairwise disjoint classes $C_0, C_1, \ldots, C_k$ is said equitable if all the classes C_i $(1 \leq i \leq k)$ have the same cardinality.

The exceptional set C_0 (which may be empty) has only a technical purpose: it makes it possible that all other classes have exactly the same number of vertices.

An equitable partition $C_0, C_1, \ldots, C_k$, with C_0 being the exceptional set, is called ε-regular if $|C_0| < \varepsilon|V|$ and all but at most εk^2 of the pairs (C_i, C_j) are ε-regular $(1 \leq i < j \leq k)$.

Note: A singleton partition is ε-regular for every value of ε.

Szemerédi's Lemma

Theorem (Szemerédi, 1976) *For every $\varepsilon > 0$ and every positive integer t there is an integer $Q = Q(\varepsilon, t)$ such that every graph with $n > Q$ vertices has an ε-regular partition into $k + 1$ classes, where $t \leq k \leq Q$.*

Notes:

Theorem is trivial for sparse graphs (but not for dense ones).

A large value of t ensures that the C_i's are sufficiently small.

The upper bound Q on the number of partitions guarantees that for large graphs the partition sets are large too.

Talk's Outline

- Szemerédi's regularity lemma
- Finding regular partitions in polynomial time
- The regularity lemma and pairwise clustering
- Experimental results

Algorithmic Issues

Theorem (Alon et al., FOCS 1992) *For every fixed $\varepsilon > 0$ and $t \geq 1$ a Szemerédi partition con be found in $O(M(n))$ sequential time, where $M(n) = O(n^{2.376})$ is the time for multiplying two $n \times n$ matrices with 0/1 entries over the integers.*

The partition can be found in time $O(\log n)$ on a EREW PRAM with a polynomial number of parallel processors.

Checking Regularity

Lemma (Alon et al., 1992) *Let H be a bipartite graph with equal classes $|A| = |B| = n$. Let $2n^{-1/4} < \varepsilon < \frac{1}{16}$.*

There is an $O(n^{2.376})$ algorithm that verifies that H is ε-regular or finds two subsets $A' \subseteq A$, $B' \subseteq B$, $|A'| \geq \frac{\varepsilon^4}{4}n$, $|B'| \geq \frac{\varepsilon^4}{4}n$ and $|d(A', B') - d(A, B)| \geq \varepsilon^4$.

Index of Partition

Given an equitable partition P of a graph $G = (V, E)$ into classes $C_0, C_1 \ldots C_k$, the index of partition is defined as follows:

$$ind(P) = \frac{1}{k^2} \sum_{s=1}^{k} \sum_{t=s+1}^{k} d(C_s, C_t)^2. \tag{4}$$

Since $0 \leq d(C_s, C_t) \leq 1$, $1 \leq s, t, \leq k$, we have $ind(P) \leq \frac{1}{2}$.

Refining a Non-regular Partition

Lemma (Szemerédi, 1976) *Let $G = (V, E)$ be a graph with n vertices. Let P be an equitable partition of V into classes $C_0, C_1 \ldots C_k$, with C_0 being the exceptional class.*
Let $\gamma > 0$. Let k be the least positive integer such that $4^k > 600\gamma^{-5}$.

If more than γk^2 pairs (C_s, C_t) $(1 \leq s < t \leq k)$ are γ-irregular, then there is an equitable partition Q of V into $1 + k4^k$ classes, the cardinality of the exceptional class being at most

$$|C_0| + \frac{n}{4^k}$$

and such that

$$ind(Q) > ind(P) + \frac{\gamma^5}{20}.$$

Alon et al.s Algorithm: Initialization

Given any $\varepsilon > 0$ and a positive integer t, let b be the least positive integer such that

$$4^b > 600\Big(\frac{\varepsilon^4}{16}\Big)^{-5}, \quad b \geq t.$$

Given a graph $G = (V, E)$ a Szemerédi partition can be constructed using the following $O(M(n)) = O(n^{2.376})$ algorithm.

Alon et al.'s Algorithm: Main Loop

1. **Create the initial partition**: arbitrarily divide the set V into an equitable partition P_1 with classes $C_0, C_1, \ldots, C_b$ where $|C_i| = \lfloor n/b \rfloor$, $i = 1 \ldots b$ and $|C_0| < b$. Let $k_1 = b$.

2. **Check regularity**: for every pair C_r, C_s of P_i verify if it is ε-regular or find $X \subseteq C_r$, $Y \subseteq C_s$, $|X| \geq \frac{\varepsilon^4}{16}|C_r|$, $|Y| \geq \frac{\varepsilon^4}{16}|C_s|$ such that

$$|d(X,Y) - d(C_s,C_r)| \geq \varepsilon^4. \tag{5}$$

3. **Count regular pairs**: if there are at most $\varepsilon\binom{k_i}{2}$ pairs that are not verified as ε-regular, then halt. P_i is an ε-regular partition.

4. **Refine**: apply the Lemma of Szemerédi, where $P = P_i$, $k = k_i$, $\gamma = \frac{\varepsilon^4}{16}$ and obtain a partition P' with $1 + k_i 4^{k_i}$ classes.

5. Let $k_{i+1} = k_i 4^{k_i}$, $P_{i+1} = P'$, $i = i + 1$ and go to step 2.

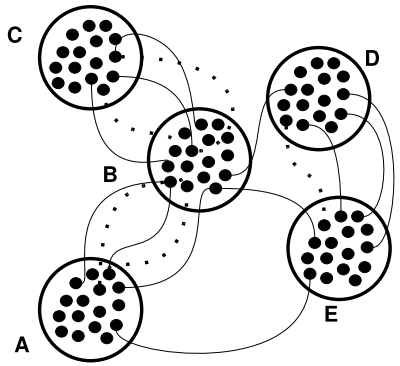
Talk's Outline

- Szemerédi's regularity lemma
- Finding regular partitions in polynomial time
- The regularity lemma and pairwise clustering
- Experimental results

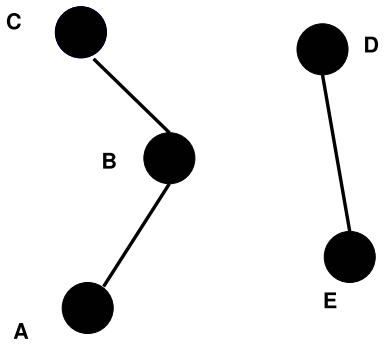
Reduced Graphs

Given a graph $G = (V, E)$, a partition P of the vertex-set V into the sets $C_1, C_2, \ldots C_k$ and two parameters ε and d, the reduced graph R is defined as follows.

The vertices of R are the clusters $C_1, C_2, \ldots C_k$, and C_i is adjacent to C_j if (C_i, C_j) is ε-regular with density more than d.



(a)



(b)

Expanded Graphs

Consider now a graph R and an integer t.

Let $R(t)$ be the graph obtained form R by replacing each vertex $x \in V(R)$ by a set V_x of t independent vertices, and joining $u \in V_x$ to $v \in V_y$ if and only if (x, y) is an edge in R.

$R(t)$ is a graph in which every edge of R is replaced by a copy of the complete bipartite graph K_{tt}.

The Key Lemma

Lemma (Komlós and Simonovits, 1996) *Given $d > \varepsilon > 0$, a graph R and a positive integer m, construct a graph G as follows:*

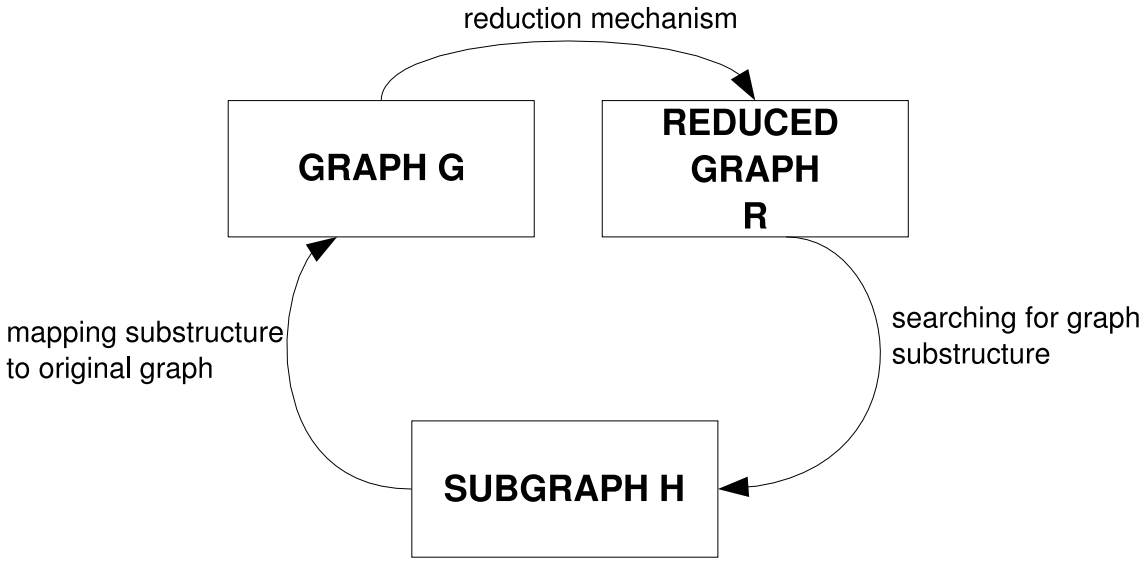
1. replace every vertex of R by m vertices
2. replace every edge of R with an ε-regular pair of density at least d.

Let H be a subgraph of $R(t)$ with h vertices and maximum degree $\Delta > 0$, and let $\varepsilon_0 = (d-\varepsilon)^{\Delta}/(2+\Delta)$.

If $\varepsilon \leq \varepsilon_0$ and $t - 1 \leq \varepsilon_0 m$, then H is embeddable into G (i.e., G contains a subgraph isomorphic to H).

Using the Key Lemma

A direct consequence of the Key Lemma is that it is possible to search for significant substructures in a reduced graph R in order to find common subgraphs of R and the original graph.



Using Weighted Graphs

Simple verifications on the previous unweighted definitions show that the algorithms are not influenced by edge-weights (cfr. Czygrinow and Rödl, *SIAM J Comp.* 2000).

In this case, the density between the sets of a pair (A, B) in a becomes:

$$d = d_\omega(A, B) = \frac{\sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \omega(a_i, b_j)}{|A||B|}$$

which is the weighted counterpart of (1).

Application to Pairwise Clustering

Motivated by the previous discussion, here we propose a two-phase clustering strategy.

In the first phase, we apply the regularity lemma in order to build a compact, yet informative representation of the original data set, in the form of a reduced graph.

In the second phase, a pairwise clustering method is used to find meaningful structures in the reduced graph.

In our experiments we used the dominant set approach (Pavan and Pelillo, *CVPR* 2003, *PAMI* 2007).

Note: Our approach differs from out-of-sample methods (e.g., Fowlkes et al., *PAMI*, 2004; Pavan and Pelillo *NIPS*, 2004).

Dealing with Intra-class Similarities

To take into account intra-class similarities we introduced in the partitioning algorithm a rule to select elements to be dispatched to new subsets.

Specifically, the criterion used is the average weighted degree

$$\text{awdeg}_S(i) = \frac{1}{|S|} \sum_{j \in S} \omega(i,j) \qquad S \subseteq V.$$

All elements in the current subset S are listed in decreasing order by average weighted degree.

The partition of S takes place simply by subdividing the ordered sequence of elements into the desired number of subsets.

Talk's Outline

- Szemerédi's regularity lemma
- Finding regular partitions in polynomial time
- The regularity lemma and pairwise clustering
- Experimental results

Implementation Details

We stop the algorithm either when a regular partition has been found or when the subset size becomes smaller than a predetermined threshold.

The next-iteration number of subsets of the original procedure is intractable, and we therefore decided to split every subset, from an iteration to the next one, using a (typically small) user-defined parameter.

We assign the elements of the exceptional set to the closest cluster according to a predefined distance measure.

Experiments on UCI Datasets

We selected the following (two-class) datasets:

- Johns Hopkins University Ionosphere database (352 elements)
- Haberman's Survival database (306 elements)
- Pima Indians Diabetes database (768 elements).

The similarity between data items was computed as

$$w(i,j) = \exp(-||\mathbf{v}_i - \mathbf{v}_j||^2/\sigma^2)$$

where $\mathbf{v}_i$ is the i-th vector of the dataset and σ is a positive real number which affects the decreasing rate of w.

Results on UCI Datasets

Dataset	Size	Classif. Accuracy		Speedup	R.G. size	Compression rate
		Two-phase	Plain DS			
Ionosphere	351	72%	67%	1.13	4	98.9%
Haberman	306	74%	73%	2.16	128	58.1%
Pima	768	65%	65%	2.45	256	66.7%

Image Segmentation

Our two-phase strategy was also applied to the problem of segmenting brightness images.

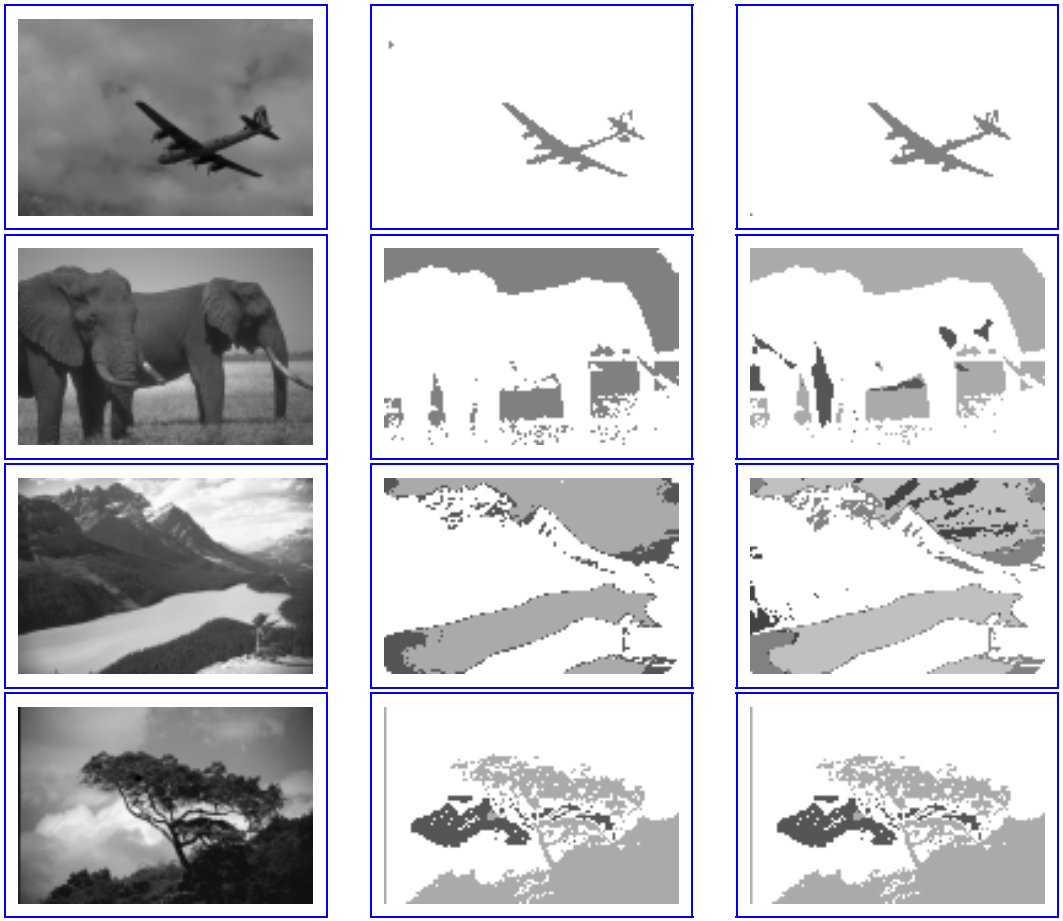
Here, the image is represented as an edge-weighted graph where vertices represent pixels and edge-weights reflect the similarity between pixels.

Similarity between pixels i and j was measured by

$$w(i,j) = \exp\{-(I(i) - I(j))^2/\sigma^2\}$$

where $I(i)$ is the normalized intensity value at node i.

Image Segmentation



Segmentation Results

Image	Pixels	R.G. size	Compression rate	Speedup
Airplane	9600	128(75)	98.7%	4.23
Elephants	9600	32(300)	99.7%	17.79
Lake	9600	128(75)	98.7%	20.94
Tree	9600	64(150)	99.3%	16.63

The table summarizes information concerning the image segmentation experiments. Each row represents an image, while the columns represent (left to right): the number of pixels in the original image, the number of vertices in the reduced graph (with the subset dimension in parenthesis), the compression rate, and the speedup achieved using our approach w.r.t. plain dominant-set.

Conclusions

- Imported Szemerédi's Regularity Lemma into CV/PR/ML fields
- Applications to pairwise clustering as a graph compression device
- Preliminary results encouraging

Future Work

- More extensive experimentation
- Alternative partitioning algorithms
  (e.g., Frieze and Kannan, *EJC* 1999)
- Extension to hypergraphs
  (Czygrinow and Rödl, *SIAM J Comp.* 2000).
- Other applications in CV/PR/ML fields?