A Major Qualifying Project Report
ON

Experimental Improvements to Regularity Clustering

Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE

In Partial Fulfillment of the Requirement for the
Degree of Bachelor of Science

by

Keleigh O'Neil
Stephen L. Peters

UNDER THE GUIDANCE OF
Professor Peter R. Christopher
Professor Gábor N. Sárközy

February 23, 2014

Abstract

Data clustering is an immensely powerful tool. The analysis of big data has led to many clustering techniques. Among these techniques is Regularity Clustering, a new technique based on Abel Prize winner Endre Szemerédi's Regularity Lemma. Regularity Clustering has been shown to outperform industry standard clustering techniques in many circumstances. In this report we present new methods of executing Regularity Clustering. Among these methods one, which we call the most recurring construction method, outperforms the standard Regularity Clustering method by a significant margin. We also present empirical evidence indicating when Regularity Clustering performs well.

Acknowledgements

First and foremost, we would like to thank our advisors Professor Peter Christopher and Professor Gábor Sárközy. Together their guidance has proved invaluable in navigating the complexities of our research. We truly could not have had the success we did without them. We would also like to thank Fei Song and Shubhendu Trivedi for their assistance in managing our implementation of Regularity Clustering. Their expertise of the implementation saved us countless hours and improved our ability to perform our research. Finally we would like to thank Professors Stanley Selkow and Neil Heffernan for their continued interest in our work.

Contents

List of Figures

Chapter 1

Introduction

Big Data has become a major topic in recent years as the amount of information has increased exponentially along with technological advancements. As we are able to store more and more data it becomes a question of what we can learn from this data. One idea is to try to organize the data together so that the data points that are grouped together share common attributes. This way when a new datum point is introduced, if one can accurately predict the group to which the datum point belongs then one could also predict the value of unknown attributes that the group shares. This is the essence of data clustering, to predict an attribute of new data based on the values of old data. This tool is immensely effective in answering some very important questions: What kind of skin disease does one have? Is the tumor one just found malignant or benign? Will the surgery the doctor is recommending add five years to a patient's life? But data clustering is not limited to questions like these, we can also predict how well a student will learn from a particular tutoring technique, or how many miles per gallon a car gets. Any question for which we have data to compare can be answered with relatively good accuracy using data clustering techniques.

Many clustering techniques have been created which group data based on similar characteristics. Some of these methods include spectral clustering, k-means clustering, density-based clustering, and probabilistic clustering [4]. A new and promising type of clustering, Regularity Clustering, was recently introduced [16].

In 1975 in his proof of the celebrated Szemerédi's Theorem [18], Endre Szemerédi proved what is known today as Szemerédi's Regularity Lemma, which has turned out to be an extremely powerful result in mathematics and

theoretical computer science. The Regularity Lemma is applicable to many problems across combinatorics and extremal graph theory, such as Ramsey-Turan theory [8] [18] [19], the (6,3) extremal hypergraph problem [15], and its applications with large forbidden graphs [5] to name a few. The lemma has been used to prove some of the intricate conjectures of the last 30 years. This result is so important that Szemerédi was awarded the Abel Prize, the unofficial Nobel Prize in Mathematics, for this work in 2012 [14].

A major criticism of the Regularity Lemma comes from its inability to be used in real world applications as it only worked on graphs that are so large they could not possibly be represented. In fact, Field's Medal winner Sir Timothy Gowers writes in his paper about Szemerédi that "the theorem (is) well beyond the realms of any practical applications" [11]. Until recently the Regularity Lemma was considered a purely theoretical result. However in 2012 Sárközy, Song, Szemerédi, and Trivedi made a modified version of the algorithm used to prove the Regularity Lemma. While it has not been proven that this modified algorithm will ever produce the results of the original Regularity Lemma, the size requirement for the input to the modified algorithm is practical [16].

The idea was to use this modified Regularity Lemma algorithm, in conjunction with modern data clustering techniques to produce a new clustering technique termed Regularity Clustering. Their results were very promising, despite the lack of understanding about the theory behind the modified algorithm. In this paper we build on the results of Sárközy et al. and improve the accuracy of Regularity Clustering while also classifying some of the attributes that make Regularity Clustering effective.

We created six new variations of Regularity Clustering each of which performed better than the variation created by Sárközy et al. One variation in particular, which we call the most recurring construction, significantly outperformed the standard variation. We also found evidence that the success of Regularity Clustering is influenced by the ratio of data points to target clusters of the dataset. This discovery was groundbreaking as previously there was no way to predict if Regularity Clustering would perform well on any given dataset. With this discovery we are one step closer to being able to confidently use Regularity Clustering for real world applications. Used to their full effectiveness these improvements have the potential to provide the means to answering very difficult questions, to improve the quality of life of people around the world by improving individualized care and education, and even to save lives by providing quick and accurate diagnosis of illness.

Chapter 2

Background

In this chapter we present the Regularity Lemma, an outline of its proof, algorithmic versions of the lemma, data clustering, and we show how the Regularity Lemma can be used to improve upon modern clustering techniques. First we will cover definitions and concepts that are vital to the proper understanding of these topics. For the rest of this chapter, let $G = (V, E)$ be a graph where V is the set of vertices of the graph G and E is the set of edges of the graph G.

2.1 Important Concepts

The purpose of the Regularity Lemma is to partition the vertices of a graph into classes that behave almost randomly with each other. The concepts required to discuss the Regularity Lemma include density, ε-regular pairs, ε-regular partitions, refinements of partitions, and the index of a partition.

Definition 1. *For disjoint subsets of vertices A and B, the **density** of the pair, denoted $d(A, B)$ is the ratio of edges between the pair to the maximum possible number of edges between subsets of this size. This is equal to the number of edges between A and B, denoted $\| A, B \|$, divided by the product of the size of A and the size of B. Thus:*

$$d(A, B) = \frac{\| A, B \|}{|A||B|} \tag{2.1}$$

For example, consider the bipartite graph depicted in Figure 2.1. 9 of the possible 25 edges are present; therefore the density of the graph is $\frac{9}{25} = 0.36$.

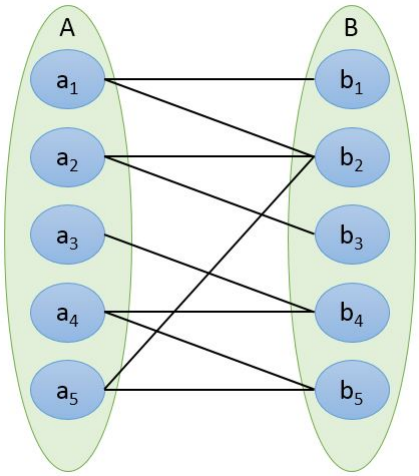

Figure 2.1: Example graph for density.

Definition 2. *A pair of disjoint subsets A and B of V is* ***ε-regular*** *for some $\varepsilon > 0$ if for every subset X of A and subset Y of B which are sufficiently large ($|X| \geq \varepsilon|A|$ and $|Y| \geq \varepsilon|B|$), the density of the pair X, Y differs from the density of the pair A, B by at most ε. That is:*

$$|d(A, B) - d(X, Y)| \leq \varepsilon. \tag{2.2}$$

If the edges between A and B were distributed randomly we would expect to observe this behavior; therefore we can think of the edges of an ε-regular pair as being distributed ε close to randomly.

As an example, let $\varepsilon = 0.25$ and consider the graph in Figure 2.1. Since $|A| = |B| = 5$ we must check for all pairs X and Y where $|X| \geq \varepsilon|A| = 1.25$ and $|Y| \geq \varepsilon|B| = 1.25$ that it has density greater than $0.36 - \varepsilon = 0.11$ or less than $0.36 + \varepsilon = 0.61$. Since there are 26 subsets of A larger than 1.25 we have 676 pairs to consider.

Consider the pairs of square vertices in Figure 2.2. The pair on the left is composed of subsets that are large enough, yet the density of the pair is 0.75. This is enough to show that the pair is ε-irregular. The center pair's density is 0 and thus this pair also shows ε-irregularity. Finally the pair on the right has density $\frac{5}{9}$ which falls within our bounds for ε-regularity.

Definition 3. *A partition $P(V) = V_0 \cup V_1 \cup V_2 \cup ... \cup V_k$ of the vertices of a*

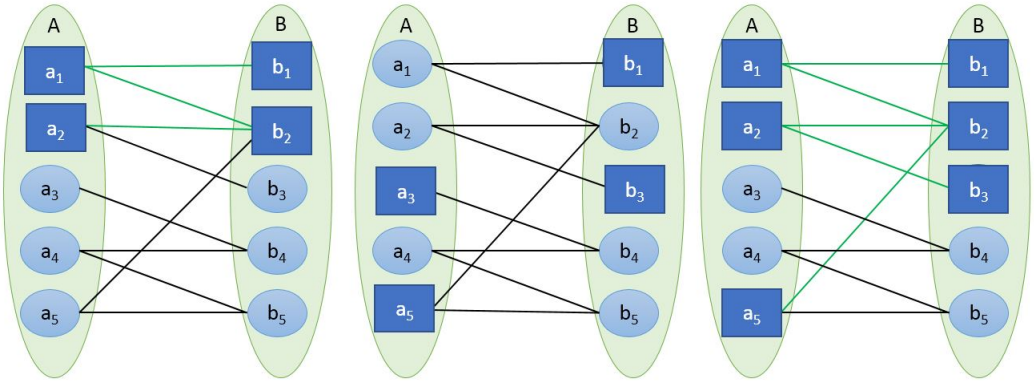

Figure 2.2: ε-regular pair examples.

graph is called an ϵ***-regular partition*** *of* G *if all but at most* εk^2 *of the pairs of sets in the partition* (V_i, V_j) *form an* ε*-regular pair in* G *where* k *is the number of non-exceptional classes in the partition (*V_0 *is the exceptional set where* $V_1, ..., V_k$ *are non-exceptional). Otherwise it is an* ε*-irregular partition.*

Definition 4. *A partition* Q *of the set* S *is considered a* ***refinement*** *of a partition* P *of* S *if every element of* Q *is a subset of some element of* P*. That is if* $Q = \{Q_1, Q_2, ..., Q_j\}$ *and* $P = \{P_1, P_2, ..., P_k\}$ *then each* Q_t *in* Q *is a subset of some* P_s *in* P*. In this case we say that* Q *is finer than* P *and* P *is courser than* Q*.*

When we consider the refining of an ε-(ir)regular partition we usually do not consider the exceptional set. That is, the exceptional set of the refinement need not be a subset of any set from the original ε-(ir)regular partition.

Definition 5. *The* ***index*** *of a partition is the sum of the squares of the densities of every pair in the partition divided by (about) twice the number of pairs. More precisely:*

$$q(P) = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} d^2(X_i, X_j). \tag{2.3}$$

Since the square of the density of a pair is at most one, the sum of these squares is at most the number of pairs ($\binom{k}{2}$ or $\frac{k(k-1)}{2}$) thus the index of a

partition is bounded above by $\frac{1}{2}$. The index of a partition is closely related to its ε-regularity and will give us a notion of how close a partition is to being ε-regular.

Now we are able to state the Regularity Lemma:

Theorem 1 (Szemerédi [20] see also in [7])**.** *For every $\varepsilon > 0$ and m there exist two integers $M(\varepsilon, m)$ and $N(\varepsilon, m)$ such that for every graph with $n \geq N(\varepsilon, m)$ vertices there exists a partition of the vertex $P(V) = \{V_0, V_1, V_2, ..., V_k\}$ set into $k+1$ disjoint subsets with the following properties:*

1. *$m \leq k \leq M(\varepsilon, m)$ where $k+1$ is the number of classes in our partition (k normal partition classes plus the exceptional set).*
2. *The exceptional set V_0 has size less than or equal to ε times the order of the graph ($|V_0| \leq \varepsilon n$).*
3. *Each subset in the partition has the same cardinality excluding the exceptional set. $|V_1| = |V_2| = ... = |V_k|$*
4. *Fewer than εk^2 of the pairs are ε-irregular.*

As we defined earlier, a partition that meets these requirements is called an ε-regular partition of the vertices of the graph. A partition that fails only the fourth requirement is called an ε-irregular partition.

2.2 Outline of the Proof

We provide the reader with an outline of the proof of the Regularity Lemma because it provides insight into the decisions that must be made when implementing a Regularity Clustering algorithm. Readers interested in the proof this outline is modeled on are referred to [7]. To prove the Regularity Lemma we employ four lemmas.

Lemma 1. *If $\boldsymbol{C}$ is a partition of C and $\boldsymbol{D}$ is a partition of D, then $q(\boldsymbol{C}, \boldsymbol{D}) \geq q(C, D)$ where*

$$q(C, D) = \frac{d^2(C, D)}{k^2} \tag{2.4}$$

and

$$q(\boldsymbol{C}, \boldsymbol{D}) = q(\boldsymbol{C}) + q(\boldsymbol{D}) + \sum_{i=1}^{|C|} \sum_{j=1}^{|D|} q(C_i, D_j). \tag{2.5}$$

In other words, partitioning a pair cannot make the index less. With Lemma 1 we can tackle the next lemma which states the following:

Lemma 2. *If P and P' are partitions of V and if P' refines P, then $q(P') \geq q(P)$.*

This follows directly from the definition of a refinement of a partition and repeated applications of Lemma 1.

Lemma 3. *Let (C, D) be an ε-irregular pair. If (C', D') is a witness of ε-irregularity (subsets of C and D which show that C and D are ε-irregular) where $C' \subseteq C$ and $D' \subseteq D$, then partitioning C into $C^* = \{C', C \backslash C'\}$ and D into $D^* = \{D', D \backslash D'\}$ guarantees $q(C^*, D^*) > q(C, D)$.*

That is, separating an ε-irregular witness is guaranteed to increase the index.

Lemma 4. *An ε-irregular partition can be refined in such a way that the index increases by at least a constant amount $(\frac{\varepsilon^5}{2})$.*

Lemma 4 follows from Lemma 3 and the fact that an ε-irregular partition contains at least εk^2 irregular pairs.

We call the refinement guaranteed by Lemma 4 an **intermediate refinement**. Recall that the index is bounded above by $\frac{1}{2}$. This means that there is an upper bound $(\frac{2}{\varepsilon^5})$ on the number of times this refinement can be applied before it must be the case that the result has less than εk^2 irregular pairs. Additionally, this partition can be further refined (without decreasing the index, by Lemma 2) into much smaller, but equally sized parts, where the leftover vertices are added to the exceptional set. Further, we can choose this size small enough to guarantee that the number of vertices added to the exceptional set is not too large. We call this partition the **iteration's partition**.

The intermediate partition described above is achieved by taking the unique maximal partition that refines every ε-irregular witness. For example, if a partition class is ε-irregular with three other partition classes and each witness intersects each other witness, the refinement of this piece will have $2^3 = 8$ pieces. Figure 2.3 depicts what is occurring, where A, B, and C represent the ε-irregular witnesses and each color, including the white section that does not belong to A, B or C, represents a class in the intermediate refinement.

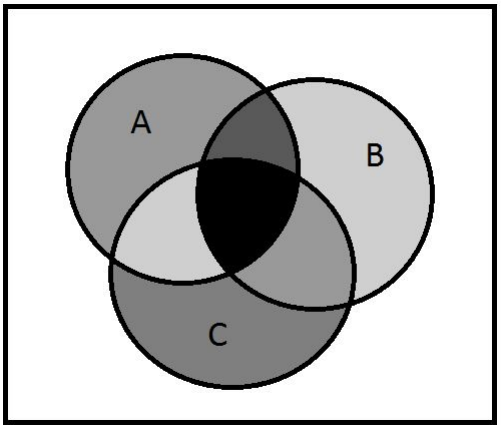

Figure 2.3: Venn-Diagram refinement example.

From the intermediate partition we construct the iteration's partition by dividing each element of the intermediate partition into a maximal number of pieces of size $\frac{c}{4^k}$ where c is the size of the classes of our original equitable partition. Since each of the k classes of our original partition are divided into a maximum of 2^k classes and from each we could add a maximum of $\frac{c}{4^k}$ vertices to V_0, we have a maximum of $\frac{ck2^k}{4^k} = \frac{n}{2^k}$ vertices added to the exceptional set (where n is the number of vertices in the graph). It is important to note that the iteration's refinement has $k4^k$ pieces and thus has exponentially more classes of much smaller size than the original.

Knowing Lemma 4, it is possible to obtain a partition that is guaranteed by the Regularity Lemma. At each step in the partitioning we are going to apply Lemma 4 to our current ε-irregular partition. This will yield a new partition with an index at least $\frac{\varepsilon^5}{2}$ higher than the previous partition. Repeated application of this must yield an ε-regular partition as the index is bounded above by $\frac{1}{2}$. Thus we have an upper bound on the number of iterations ($\frac{2}{\varepsilon^5}$) the partitioning can take before regularity is achieved. During each iteration, the size of the exceptional set grows by at most $\frac{n}{2^k}$, thus over the course of the partitioning the exceptional set will grow by at most $\frac{n}{2^{k-1}\varepsilon^5}$. All that remains is to choose the parameters of our initial partition to ensure we do not exceed our bounds. We must choose k (size of the initial partition) that is large enough that after $\frac{2}{\varepsilon^5}$ iterations so that the exceptional set does not grow more than $\frac{\varepsilon n}{2}$. Thus we choose k such that $2^{k-1} \geq \frac{2}{\varepsilon^6}$

which is equivalent to $k \geq 2 - log_2(\varepsilon^6)$. We then choose M, the upper bound on the number of sets in the partition, which grows from x to $x4^x$ each iteration, to be $f^{\frac{2}{\varepsilon^5}}(k)$ (applying f to k $\frac{2}{\varepsilon^5}$ times) where $f(x) = x4^x$. Graphs of order less than M are trivially partitioned into sets of size one producing an ε-regular partition. For any graph of order larger than M this partitioning produces a non-trivial ε-regular partition of V. To give some reference on the size of these number for $\varepsilon = 0.93, k = 3$ and $M = 126,021$ and for $\varepsilon = 0.92, k = 3$ and $M = 15,880,788,357$. As you can see, the tower function that defines M increases exceptionally fast as ε decreases. It increases so fast that graphs large enough to guarantee an ε-regular partition for even $\varepsilon = 0.5$ (considered quite large) are so large it would not be feasible to represent one in practice. In 1998, W.T. Gowers proved the tower function lower bound is necessary for the Regularity Lemma to work on all graphs [10]. This was done by constructing an extremely degenerate example that does not have an ε-regular partition until the size of the graph surpasses the tower function lower bound.

2.2.1 Refinement Example

The following example is designed purely to explain the refinement process. There is no underlying graph and we are using a different notion of ε-regularity

Let the following be defined:

$\varepsilon = \frac{1}{6}$, the set to be partitioned $V = \{1, 2, ..., 30\}$, the initial partition $P = \{P_0, P_1, P_2, P_3, P_4\}$, where

$$\text{the exceptional set} P_0 = \{1, 23\},$$
$$P_1 = \{2, 6, 12, 17, 24, 28, 30\},$$
$$P_2 = \{3, 7, 14, 19, 21, 22, 29\},$$
$$P_3 = \{4, 5, 8, 11, 13, 18, 27\},$$
$$P_4 = \{9, 10, 15, 16, 20, 25, 26\}.$$

A pair (A, B) is called an ε-irregular witness here if $|A| \geq \varepsilon|P_x|, |B| \geq \varepsilon|P_x|$, and A and B are composed only of prime numbers. When we examine the pairs for this example we find the following witnesses based on this new definition.

The pair (P_1, P_2) yields witness ({2, 17}, {3, 7, 29}).

The pair (P_1, P_3) yields witness ($\{2, 17\}$, $\{5, 13\}$).
The pair (P_1, P_4) is a regular pair.
The pair (P_2, P_3) yields witness ($\{3, 7\}$, $\{5, 11\}$).
The pair (P_3, P_4) is a regular pair.

Now we construct P_{ij} for $1 \leq i, j \leq 4$ where P_{ij} is the witness or P_i induced by P_j unioned with the complement of the witness. So:

$$P_{12} = \{\{2, 17\}, \{6, 12, 24, 28, 30\}\}$$
$$P_{13} = \{\{2, 17\}, \{6, 12, 24, 28, 30\}\}$$
$$P_{14} = \{\{2, 6, 12, 17, 24, 28, 30\}\}$$
$$P_{21} = \{\{3, 7, 29\}, \{14, 19, 21, 22\}\}$$
$$P_{23} = \{\{3, 7\}, \{14, 19, 21, 22, 29\}\}$$
$$P_{24} = \{\{3, 7, 14, 19, 21, 22, 29\}\}$$
$$P_{31} = \{\{5, 13\}, \{4, 8, 17, 27, 11\}\}$$
$$P_{32} = \{\{5, 11\}, \{4, 8, 13, 18, 27\}\}$$
$$P_{34} = \{\{4, 5, 8, 11, 13, 18, 27\}\}$$
$$P_{41} = \{\{9, 10, 15, 16, 20, 25, 26\}\}$$
$$P_{42} = \{\{9, 10, 15, 16, 20, 25, 26\}\}$$
$$P_{43} = \{\{9, 10, 15, 16, 20, 25, 26\}\}$$

Now we construct P_i^* for each $1 \leq x \leq 4$ where P_i^* is the unique minimal partition that refines each of P_{ij} so

$$P_1^* = \{\{2, 17\}, \{6, 12, 24, 28, 30\}\}$$
$$P_2^* = \{\{3, 7\}, \{29\}, \{14, 19, 21, 22\}\}$$
$$P_3^* = \{\{5\}, \{11\}, \{13\}, \{4, 8, 17, 27\}\}$$
$$P_4^* = \{\{9, 10, 15, 16, 20, 25, 26\}\}$$

From these we construct the intermediate partition :

$$P^* = P_0 \cup P_1^* \cup P_2^* \cup P_3^* \cup P_4^*$$
$$P^* = \{\{1, 23\}, \{2, 17\}, \{6, 12, 24, 28, 30\}, \{3, 7\}, \{29\}, \{14, 19, 21, 22\},$$

$$\{5\}, \{11\}, \{13\}, \{4, 8, 17, 27\}, \{9, 10, 15, 16, 20, 25, 26\}\}$$

The final step in the refinement process is to reduce each class size to a size small enough that the adding the left over pieces to the exceptional size will not cause the exceptional size to grow by too much. For this example, we let that size be 2. then the iteration's partition is:

$$P' = \{\{1, 23, 30, 29, 5, 11, 13, 26\}, \{2, 17\}, \{6, 12\}, \{24, 28\}, \{3, 7\},$$
$$\{14, 19\}, \{21, 22\}, \{4, 8\}, \{17, 27\}, \{9, 10\}, \{15, 16\}, \{20, 25\}\}$$

2.3 Algorithmic Versions

The astronomical size requirements is not the only obstacle to implementing the Regularity Lemma. Lemma 4 described in the previous section requires that we identify those pairs which are ε-irregular. Naively, this process takes exponential time as we need to check every pair of subsets and the number of subsets grows exponentially with the size of the set. The issue is that it can be shown that determining whether or not a pair is ε-regular is co-NP complete [1]. Yet surprisingly there are polynomial time algorithms for finding the ε-irregular witnesses required by the Regularity Lemma. In order to see how this is possible, consider the repercussions of incorrectly reporting a pair as ε-irregular. Incorrectly reporting a pair as ε-irregular increases the count of the number of ε-irregular pairs and the amount of witnesses that must be considered in the refinement process. However, neither of these results are detrimental to the algorithm. Increasing the count of ε-irregular pairs could cause the algorithm to require an extra iteration, and an additional witness could significantly increase the work required during the refinement process, but neither of these cases will result in an incorrect process. Thus the polynomial time algorithms are achieved by reporting a pair as ε-regular or ε'-irregular for $\varepsilon' < \varepsilon$. Pairs which fall between these bounds (which are both ε-regular and ε'-irregular) could produce either result, both of which are valid and we have no control over which one occurs.

One algorithmic method for identifying ε-irregular witnesses is to use the first singular value of the adjacency matrix. Another is to examine a concept called neighborhood deviation. Both of these methods will yield a polynomial time algorithm for identifying the ε-irregular pairs required for the algorithm for the Regularity Lemma, reducing the complexity from exponential to polynomial. While this is very good news in terms of computability, the constants associated with the Regularity Lemma are still too large for

practical use, regardless of the existence of a polynomial time algorithm.

2.3.1 First Singular Value Method

The singular value method developed by Frieze and Kannan in 1998 [9] uses the first singular value of the adjacency matrix of the graph when determining regularity. We first give the terminology we will be using. For any matrix A, the **first singular value** is defined as $\sigma_1(A) = max_{|x|=|y|=1}|x^T Ay|$. Second, let X_b and X_c be disjoint subsets of the vertices of the graph and define $A_{b,c}$ as the submatrix of A containing the vertices of X_b as rows and the vertices of X_c as columns. Finally, define $W_{b,c}$ as $(A_{b,c} - D)$ where D is a matrix for which every value is the average of the values in $A_{b,c}$.

Let S be a subset of the vertices in X_b and let T be a subset of the vertices in X_c. Define x_S as the vector containing 0's and 1's such that $(x_S)_i = 1$ if $i \in S$ and $(x_S)_i = 0$ if $i \notin S$. Similarly define x_T. Using these definitions we let

$$A(S,T) = \sum_{i \in S} \sum_{j \in T} A(i,j) = x_S^T A x_T. \tag{2.6}$$

We can then see that a pair (X_b, X_c) of a partition is ε-regular if and only if $|A(S,T)| \leq \varepsilon|S||T|$ where $|S| \geq \varepsilon|X_b|$ and $|T| \geq \varepsilon|X_c|$. The following Lemma from Frieze and Kannan's 1998 paper relates this definition of ε-regularity to the first singular value to show that the first singular value can be used to determine whether a pair of subsets of a partition is ε-regular.

Lemma 5 (Frieze, Kannan [9])**.** *Let W be an $R \times C$ matrix with $|R| = p$, $|C| = q$ and $\|W\|_\infty \leq 1$ and γ be a positive real. If there exist $S \subseteq R$ and $T \subseteq C$ such that $|S| \geq \gamma p, |T| \geq \gamma q$ and $|W(S,T)| \geq \gamma|S||T|$ then $\sigma_1(W) \geq \gamma^3\sqrt{pq}$. If $\sigma_1(W) \geq \gamma\sqrt{pq}$ then there exist $S \subseteq R$ and $T \subseteq C$ such that $|S| \geq \gamma' p, |T| \geq \gamma' q$ and $|W(S,T)| \geq \gamma'|S||T|$ where $\gamma' = \frac{\gamma^3}{108}$.*

Thus computing the first singular value of each pair produces a witness of ε'-irregularity if it exists where $\varepsilon' = \frac{\varepsilon^3}{108}$. If such a witness does not exist, it reports ε-regularity. This algorithm uses the first singular value of each pair (X_b, X_c) of the partition to determine regularity and produce witnesses.

2.3.2 Neighborhood Deviation Method

Rather than checking each pair for ε-regularity, which would take exponential time, one might instead consider constructing the worst pair and checking

just that instead. One might also realize that vertices with degree differing far from the average are the most promising candidates for inclusion. Upon further consideration one might also realize that degree is not enough. Instead we need a notion of pairwise degree to guarantee that this difference from the average is present in our pair. This is the notion of neighborhood deviation, which Alon, Duke, Lefmann, Rödl, and Yuster [1] used to create their algorithmic method and is formally defined as:

$$\sigma(y_1, y_2) = |N(y_1) \cap N(y_2)| - \frac{d^2}{n} \tag{2.7}$$

Here σ is the neighborhood deviation function, y_1 and y_2 are elements of the same color class of a bipartite graph, $N(v)$ denotes the neighborhood of a vertex, d is the average degree of vertices in the graph, and n is the size of the color class. The concept of neighborhood deviation can be extended to a set of vertices as follows:

$$\sigma(Y) = \frac{\sum_{y_1, y_2 \in Y} \sigma(y_1, y_2)}{|Y|^2} \tag{2.8}$$

The following statement is shown by Alon et al [1].

Lemma 6. *Let H be a bipartite graph with color classes A and B such that $|A| = |B| = n$, let d be the average degree of the vertices in H, and $0 < \varepsilon < \frac{1}{16}$ be given. Then if there exists a Y a subset of B such that $|Y| \geq \varepsilon n$ and $\sigma(Y) \geq \frac{\varepsilon^3 n}{2}$ then one of the following occurs:*

1. *$d < \sigma^3 n$*
2. *There exists a set of more than $\frac{\varepsilon^4 n}{8}$ vertices in B whose degree differs from d by at least $\varepsilon^4 n$*
3. *There are subsets A' of A and B' of B such that $|A'| \geq \frac{\varepsilon^4 n}{4}, |B'| \geq \frac{\varepsilon^4 n}{4}$, and $|d(A', B') - d(A, B)| \geq \varepsilon^4$. That is, a witness to ε^4-irregularity.*

With this we develop an algorithm for producing a witness of ε'-irregularity or verifying that the pair is ε-regular. First we compute d equal to the average degree of the vertices in H, it can be shown that if $d \leq \varepsilon^3 n$ then H must be ε-regular and we are done. If not then we count the number of vertices in B that have a degree that differs from d by at least $\varepsilon^4 n$, if there are at least $\frac{\varepsilon^4 n}{8}$ of these then at least $\frac{\varepsilon^4 n}{16}$ deviate in the same direction. Let B' be the

set of these vertices. Then $|B'| \geq \frac{\varepsilon^4 n}{16}$ and $|d(A, B') - d(A, B)| \geq \varepsilon^4$. Thus (A, B') is an ε^4-witness.

If neither of these is true then for each y_0 in B that has degree that differs from d by less than $\varepsilon^4 n$ we find the set $B_{y_0} = \{y \in B | \sigma(y_0, y) \geq 2\varepsilon^4 n\}$, this can be done by squaring H's adjacency matrix. The proof of the statement above also proves the existence of at least one such y_0 such that $|B_{y_0}| \geq \frac{\varepsilon^4 n}{4}$ thus the pair $(N(B_{y_0}), B_{y_0})$ is a witness of ε^4-irregularity.

2.4 Regularity Clustering

The objective of data clustering is to group together data points that behave similarly. Consider that this goal is similar to the goal of ε-regular partitioning. In an ε-regular partition most of the pairs (V_i, V_j) are ε-regular pairs which means the edges between V_i and V_j are distributed randomly. That is, every vertex in V_i has probability $d(V_i, V_j)$ of having an edge with every vertex of V_j. In other words, the vertices of V_i behave similarly with vertices outside of V_i. While the goals of data clustering and ε-regular partitioning are similar there are some glaring differences. The most obvious example is the third requirement of an ε-regular partition which states that each partition piece must be of equal size. Obviously it is unreasonable to expect that data clusters would all be equal in size. However the similarity motivated Sárközy, Song, Szemerédi, and Trivedi to experiment with what they call Regularity Clustering.

The idea behind Regularity Clustering is to use an ε-regular partition of the data points, generated by the Regularity Lemma to create what is known as a reduced graph and then using traditional clustering methods (Spectral, k-means, etc.) on this reduced graph to achieve our final clusters. The reduced graph has a vertex for each class of the ε-regular partition. Two vertices of the reduced graph have an edge between them if and only if the partition pieces they are assigned to form an ε-regular pair with density greater than or equal to some small value δ. This reduced graph maintains many of the properties of the original graph while being of constant (dependent only on ε) size. For our purposes, since we are not guaranteed to achieve an ε-regular partition we simple add every edge with weight equal to the density of the pair. In theory, this reduced graph should be easier to cluster due to its reduced size, and could provide better results due to the similarities between ε-regular partitions and accurate data clusters.

The choice of traditional clustering method to be performed on the reduced graph is arbitrary. We have decided upon a spectral clustering technique developed by Ng, Jordan and Weiss [13]. We chose a spectral clustering technique due to its popularity and superior performance over other techniques. Most clustering techniques, such as k-means and expectation maximization, work by estimating specific models within the data. These methods behave very poorly when the data is organized in irregular manner, such as concentric rings. On the other hand, spectral clustering methods work by analysing the spectrum of the Graph Laplacian. This effectively projects the data to a space of smaller dimension where clusters of irregular shape are much more obvious.

There are six primary steps to spectral clustering. The first step it to project the data into $\mathbb{R}^N$. Next we define an affinity matrix A based on a Gaussian Kernel K. From the affinity matrix we construct the graphs Laplacian L and then solve the eigenvalue problem $Lv = \lambda Dv$. We then select the k eigenvectors corresponding to the k lowest eigenvalues to define a k dimensional subspace P^tLP. Finally we use another clustering technique, like k-means, to form clusters in this new subspace. This process of projecting the data into this eigenspace reveals connected but not necessarily compact groups of vertices, like concentric rings. Interested readers can find a more in depth explanation of spectral clustering in [13].

The issue that we run into when trying to utilize the Regularity Lemma is that the it requires immense graphs in order to run to completion, much larger than can be feasibly clustered. Thus for realistically sized datasets (and an appropriate ε) we cannot guarantee an ε-regular partition. We can however follow the steps of the algorithm of the proof of the Regularity Lemma, with some modifications, to produce a partition that is an approximation of an ε-regular partition. There are four main modifications to the algorithm that aim to reduce the exponential refinement that occurs during each iteration.

The first of these changes is to reduce the number of ε-irregular witnesses we use to obtain our intermediate partition. The use of every ε-irregular witness is what causes the exponential refinement, as we have to refine on every intersection. Unfortunately, if we do not refine on every ε-irregular witness we have no guarantee that we will ever reach an ε-regular partition. The second modification is to the refinement of the intermediate partition to the iteration's partition. A notion of a **refinement factor** is introduced which is the number (a usual choice being between 3 and 7) of new classes that each class of the intermediate partition will be divided into when con-

structing the iteration's partition. This is the modification that changes the refinement from exponential to constant. The third modification is to what we do with the leftover vertices. If we added them all to the exceptional set, the exceptional set would grow much too quickly. Instead, all of the leftover vertices are united to form an additional refinement class for the iteration's refinement. The second modification guarantees that these vertices will be numerous enough to create another appropriately sized piece. Finally, a modification to the stopping criteria is needed. Since the algorithm is no longer guaranteed to produce an ε-regular partition, a number is chosen, usually dependent on ε and the size of the dataset, and when the size of the partition elements is less than this number the algorithm terminates. With these modifications and the use of either of the algorithmic methods for identifying ε-irregular witnesses, the algorithm can be used on reasonably sized datasets.

Sárközy, Song, Szemerédi, and Trivedi's results are very promising; however there is still so much that is not known about Regularity Clustering. There is very little known theoretically about the method due to the modifications to the algorithm. Additionally, Sárközy et al. made several heuristic choices in their implementation of the modified algorithm. The most notable of these choices is the method for choosing which ε-irregular witness(es) to refine on and how the ε-irregular witnesses should be generated.

2.5 Conclusion

In this chapter we have discussed the concept of ε-regularity, the Regularity Lemma with an outline of its proof, algorithms used to find the ε-irregular witnesses, and Regularity Clustering. We examined the heuristic choices made by Sárközy et al. in their implementation of Regularity Clustering in an attempt to improve upon their results and learn something about the attributes of the datasets for which Regularity Clustering seems to work. In the next section we discuss how we approached this problem.

Chapter 3

Methodology

In this chapter we specify the heuristic choices of Sárközy, Song, Szemerédi, and Trivedi's implementation of Regularity Clustering and discuss the choices with which we experimented. We also discuss the pros and cons of each method as we perceive them in order to attempt to justify the experiment as well as the results. Finally, we provide descriptions of the datasets we used to test our methods.

3.1 Heuristic Choices by Sárközy et al

The results found in the paper by Sárközy et al. were generated by a version of the algorithm that used unmodified versions of both the Alon et al. and Frieze-Kannan methods for generating ε-irregular witnesses. However both of these methods were created to find any ε-irregular witness, with no consideration given to the quality of the witness. With some modifications these methods could be improved for our purposes.

The algorithm then chose one witness at random to use for the refinement. The benefit of choosing the witness at random is that it generates a random sampling and thus we can expect that the refinement will be close to uniform after repeated application. However, not all witnesses are created equal and it is certainly the case that refining some witnesses brings us closer to ε-regularity than others. The disadvantage of picking at random is that we do not know if this is a good witness to partition over.

3.2 Our Heuristic Choices

Our heuristic choices can be divided into two groups; how to choose the witness(es) to refine on, and how to generate the witnesses. The methods we tested for how to choose witnesses include: best-fit, most irregular, largest and closest to half, most overlap, maximal disjoint, and paired. The methods we tested for witness generation include a most deviant construction and a most frequent construction.

3.2.1 Choice of Witness: Best-Fit

The best-fit method for choosing ε-irregular witnesses requires the selection of the witness whose size is closest to a multiple of the target size. We know the size that we are going to make the partition elements once the refinement complete. If a witness is not selected carefully one of the resulting partition classes will most likely be constructed of both vertices in and out of the witness. To try and minimize the number of vertices that cross that boundary, we select the witness whose size is closest to a multiple of the size we will make the partition elements. By doing this we hoped to construct a refinement that isolated more irregular vertices. The disadvantage of this method is that no consideration is given to the witness' irregularity. It could be the case that a witness which is not as close to perfectly sized is much more irregular which may be better to refine.

3.2.2 Choice of Witness: Most Irregular

The method of choosing the most irregular witness requires the selection of the witness whose density varies the most from the density of the original sets. We chose this witness with the hope that by dividing the most irregular witness from the other vertices, we would be more productive with each partition. We thought that by separating the vertices that were most different from each other, in one refinement the partition would be closer to a regular partition than if we had chosen a witness that was less irregular. This method has the opposite problem to best-fit in that it pays no consideration to the size of the witness, meaning that some of these very irregular vertices are likely to be mixed back in with vertices which are regular.

3.2.3 Choice of Witness: Largest / Closest to Half

The method for choosing the largest ε-irregular witness requires the selection of the generated witness that is largest in size. When we first considered this idea we thought it would make better progress since it is refining out the largest number of vertices. We then realized that this method could have a very opposite effect from what we hoped when the size of the witness surpasses half the size of the set. In response to this issue we tried a similar idea. The closest to half method requires the selections of the witness whose size is closest to half of the size of the original sets. The purpose of this method is to separate the most vertices from each other, those that are witnesses from those that are not witnesses.

3.2.4 Generating Witnesses: Most Deviant Construction

Both the Alon et al. and Frieze-Kannan algorithms produce witnesses by selecting every vertex that fits the bill. While this is the simplest way to find any ε-irregular witness it is most likely not the best witness for our purposes. Modifying the algorithm to produce witnesses that is of high quality could prove very beneficial. Our first attempt at this was to abandon the Alon et al. and Frieze-Kannan methods in favor of a simplified algorithm which will produce a "witness" regardless of the ε-regular status of the pair. Our thought process was that the classes will be refined anyway, so instead of doing nothing we should still attempt to improve the pair. The Alon et al. method creates its witnesses the majority of the time by collecting all the vertices which all have degree differing from the average in the same direction by a certain amount. We decided to use this method to generate our witnesses. For each vertex in a class we kept a running tally of the difference of its degree within each pair and that pairs density. We then constructed two "witnesses" both of one refinement factors size. The first "witness" was composed of the vertices with the largest values in the tally and the second was composed of the vertices with the last value in the tally. This method has the benefit of creating correctly sized "witnesses" doing work at each iteration, regardless of the pairs ε-regular status. However the method has the disadvantage of not necessarily generating a witness to any ε-irregular pairs reducing our knowledge of the theory behind the algorithm even more. Additionally, the method moves even further from the Regularity

Lemma by disregarding the ε-regularity of the pairs.

3.2.5 Generating Witnesses: Most Recurring Construction

For our final modification we developed an algorithm for generating a "witness" to use in the refinement. First we used either the Alon et al. or the Frieze-Kannan algorithm to construct a witness of each ε-irregular pair. Then we counted the number of times each vertex appeared in a witness which has density higher than the average, and also the number of times each vertex appears in a witness which has density less than the average. We then construct two new "witnesses" of one refinement factors size. One of these "witnesses" will be composed of those vertices which appear in the most high density witnesses. The other "witness" will be composed of the those vertices which appear in the most low density witnesses (no vertex will be used in both). This method has the benefit of creating correctly sized "witnesses" as well as taking those vertices which are a part of as many actual witnesses as possible. However the method has the disadvantage of not necessarily generating a witness to any ε-irregular pairs reducing our knowledge of the theory behind the algorithm even more.

3.3 Testing Our Choices

It is important to note that the results from Regularity Clustering vary based on the dataset being clustered and the choice of ε and refinement factor. Thus in order to test each of our choices we compared the average accuracy of our clusters over ten trials on 25 different combinations of $\varepsilon = \{0.2, 0.3, 0.4, 0.5, 0.6\}$ and refinement factor $= \{3, 4, 5, 6, 7\}$ on 10 datasets. By examining all the combinations of each, we hope to discern patterns in the results for different combinations of these heuristic choices.

The datasets we tested our methods on include: Auto-MPG [3], Contraception Method Choice [2], Dermatology [3], Haberman's Survival [3], Red Wine and White Wine [6], Steel Plates Faults and Steel Plate Pastry Faults [17], Wisconsin Diagnostic [3], and finally Yeast [3]. All of these datasets were taken from the University of California, Irvine's repository for machine learning. This repository contains hundreds of donated datasets which are used to test new machine learning techniques, like Regularity Clustering.

Many of these datasets come with predefined clusters, which allows us to test the accuracy of the clusters that we produce. We define **accuracy** as the cost of the minimum matching as defined by the Hungarian Algorithm [12] divided by the number of data points.

3.3.1 Auto-MPG

Auto-MPG is multivariate dataset with 398 instances. Each instance corresponds to a different make and model of a car. Each instance contains 8 attributes; Number of cylinders, the displacement of those cylinders, the horsepower of the car, the weight of the car, the maximum acceleration, the year the car was made in, where the car was made, and the name of the make and model. From these attributes we attempt to predict the fuel efficiency of car in miles per gallon. The MPG values range from 9.0 to 46.6. Since the MPG values correspond to the true clusters, we rounded these values to the closest integer. Additionally, since the name of the make and model of the car does not effect the mpg of the car, we removed this attribute.

3.3.2 Contraception Method Choice

Contraception Method Choice is a multivariate dataset with 1473 instances. In this dataset each instance corresponds to a wife and husband. Each instance contains 9 attributes; wife's age, wife's education, husbands education, number of children ever born, wife's religion, wife's now working, husband's occupation, standard of living index, and media exposure. From these attribute we attempt to classify the couple's contraception method choice into one for three categories; no use, long term, or short term contraception.

3.3.3 Dermatology

Dermatology is a multivariate dataset with 366 instances. In this case each instance corresponds to a patient with an erythemato-squamous disease. The nature of these diseases makes diagnosis very difficult, most of the time a biopsy is required as the symptoms are so similar. Each instance of this dataset has 34 attributes, 12 of which are basic attributes about the patients condition, such as age, itching, and family history. The remaining 24 attributes are the results of tests on skin samples. From these attributes we attempt to predict which erythemato-squamous disease the patient had.

Some of the values in the data were missing, we removed all instances which contained missing values. Additionally, we converted the names of the diseases to integer values between 1 and 6.

3.3.4 Haberman

Haberman's Survival is a multivariate dataset with 306 instances. Each instance corresponds to a patient who has undergone surgery for breast cancer. There are 3 attributes for each patient; age of the patient, the year the operation took place, and the number of positive axillary nodes detected. From these attributes we attempt to predict whether or not the patient lived five years past their surgery.

3.3.5 Red and White Wine

The Red Wine and White Wine datasets are very similar and thus we will discuss them together here. Both sets are multivariate in nature, with the red wine set having 1599 instances and the white wine set having 4898 instances. In both cases the instances correspond to red and white variants of the Portuguese "Vinho Verde" wine respectively. The instances of both datasets contain 12 attributes; fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. In both datasets we use these attribute to predict the wine score in a blind taste test, between 0 and 10 with 0 being the lowest quality wine and 10 being the highest quality wine.

3.3.6 Steel Plates Faults and Steel Plate Pastry Faults

Steel plates faults is another multivariate dataset with 1941 instances. Each instance of the steel plates faults datasets corresponds to a fault in a steel plate. Each instance contains 27 attributes about the steel such as luminosity, thickness, type of steel, etc. From these attributes we attempt to classify the fault as 1 of 7 different types of faults; pastry, Z_scratch, K_scratch, strains, dirtiness, bumps, and other faults.

We also modified this dataset to contain only two target clusters, pastry faults and other faults. We thought that if this change made a significant difference in our results that it may provide valuable information about the types of datasets that Regularity Clustering performs well on.

3.3.7 Wisconsin Diagnostic

Wisconsin Diagnostic is a multivariate dataset with 569 instances. These instances correspond to breast cancer patients at a Wisconsin hospital. Each instance has 10 attributes about the physical characteristics of the patients tumor. These attributes include; radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. From these attributes we attempt to predict the benign/malignant nature of the tumor.

3.3.8 Yeast

Yeast is a multivariate dataset with 1484 instances. These instances correspond to yeast colonies. Each instance has 8 attributes each of which is a score on a particular test for certain attributes of yeast. For example, one of the attributes is the yeast's score in the ALOM membrane spanning region prediction program. From these scores we attempt to predict the yeast's localization site. There are ten possible localization sites; cytosolic or cytoskeletal, nuclear, mitochondrial, membrane protein with no N-terminal signal, membrane protein with an uncleaved signal, membrane protein with a cleaved signal, extracellular, vacuolar, peroxisoma, and endoplasmic reticulum lumen.

3.4 Conclusion

In this chapter we described each of the heuristic choices we tested, how we planned to go about testing them, and which datasets we tested them on. In Chapter 4 we go over the results and analysis of our findings.

Chapter 4

Results And Analysis

In this chapter we present the results of our experiments to the reader. We also provide analysis of these results with regards to three very important questions: Which, if any, of our methods for witness selection/generation improve upon a random selection? What properties influence the best choice of refinement factor and ε? What properties of datasets determine whether or not Regularity Clustering will perform well? Finally we present the reader with hypotheses supported by our findings.

4.1 Data

As mentioned before, we compared the average accuracy of our clusters over ten trials on 25 different combinations of $\varepsilon = \{0.2, 0.3, 0.4, 0.5, 0.6\}$ and refinement factor $= \{3, 4, 5, 6, 7\}$ on ten datasets. Here we present the reader graphs depicting our data in a much more interpretable format. Each graph is titled with the dataset it is associated with and contains eight entries. Each entry consists of two parts, an average and best case. The average case is the average over all choices of ε and refinement factor while the best case reports the value of the ε and refinement factor which did best. The first seven entries correspond to our different Regularity Clustering Methods, while the final entry is the benchmark, the results of a standard (spectral) clustering technique. The vertical axis of the graph is associated with the percent accuracy of our clusters.

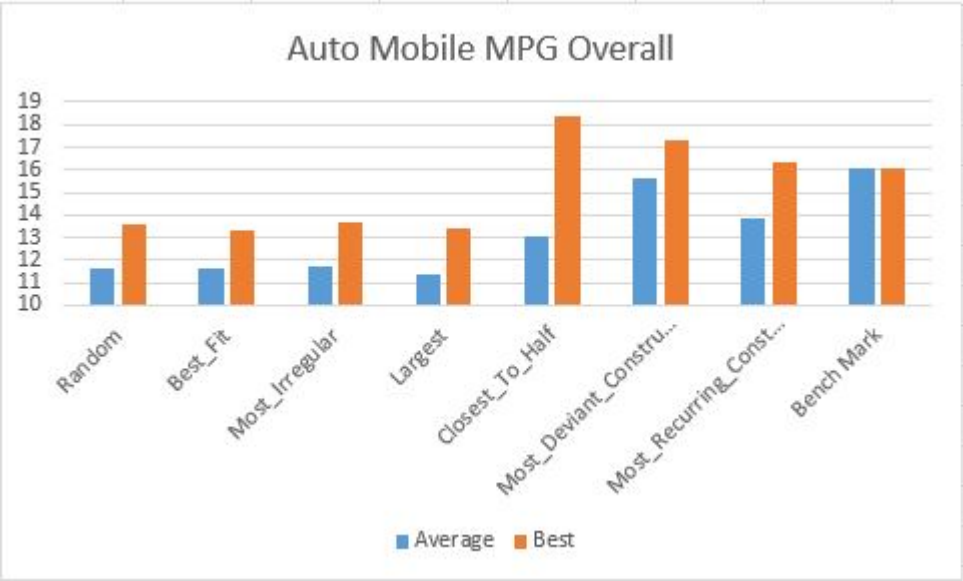

Figure 4.1: Graph of accuracy for the automobile MPG dataset.

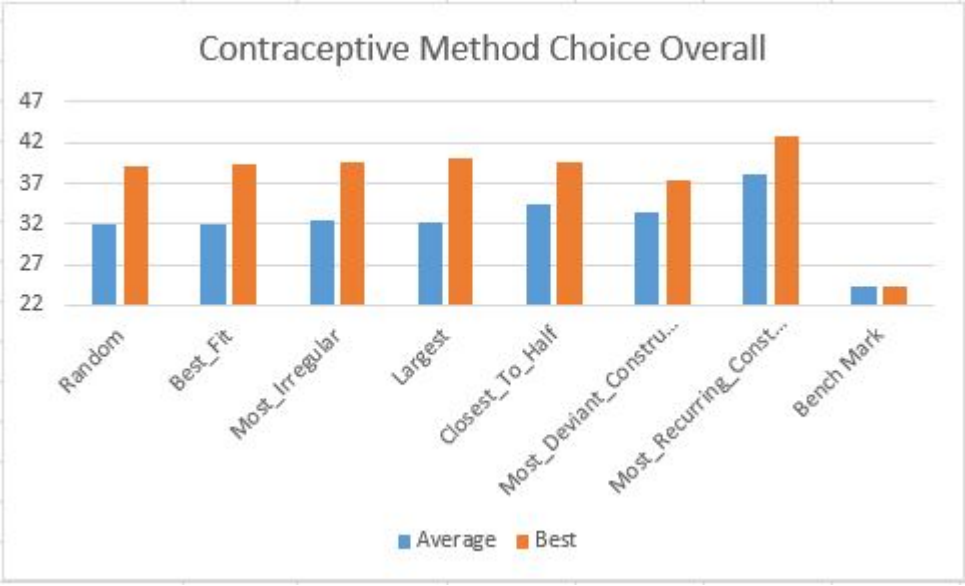

Figure 4.2: Graph of accuracy for the contraceptive method choice dataset.

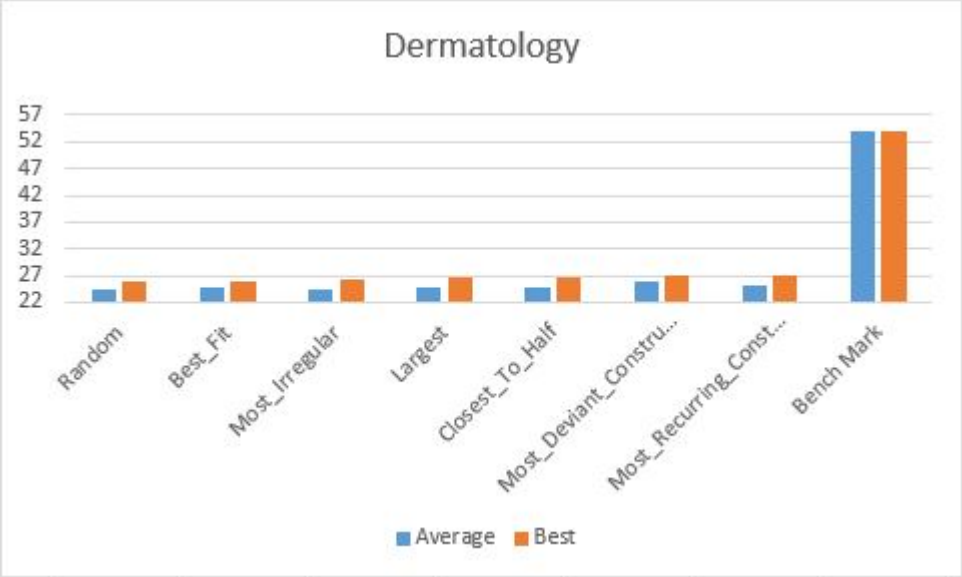

Figure 4.3: Graph of accuracy for the dermatology dataset.

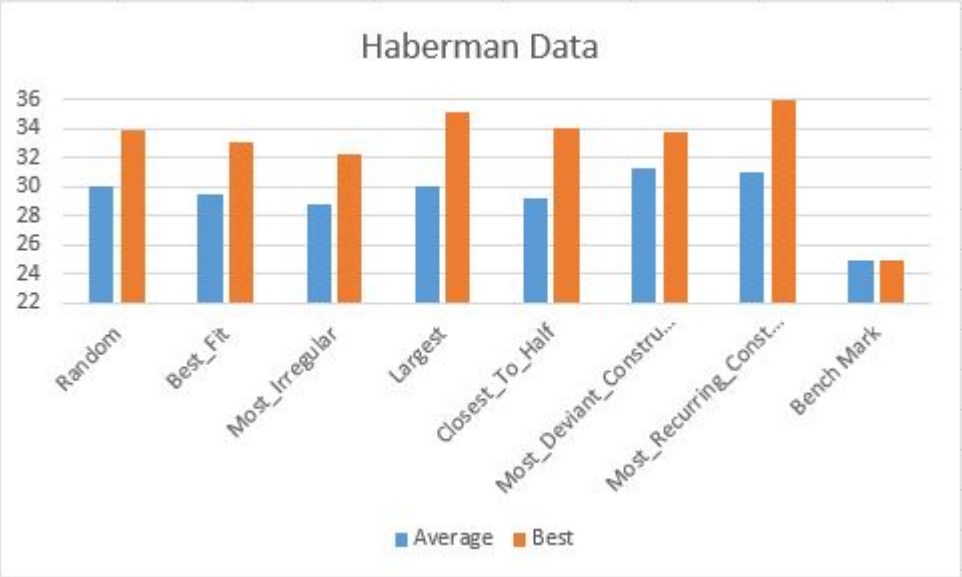

Figure 4.4: Graph of accuracy for the Haberman dataset.

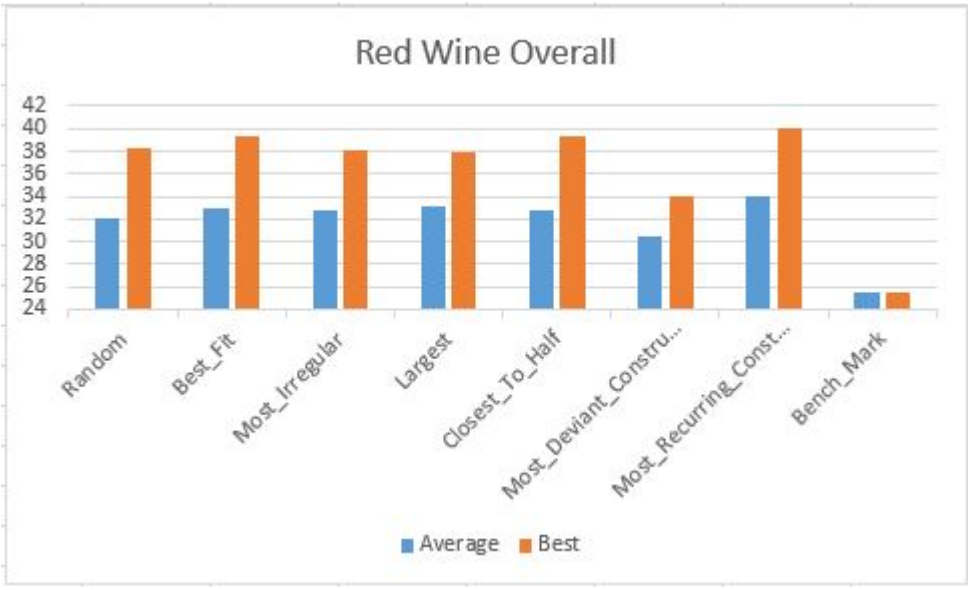

Figure 4.5: Graph of accuracy for the red wine dataset.

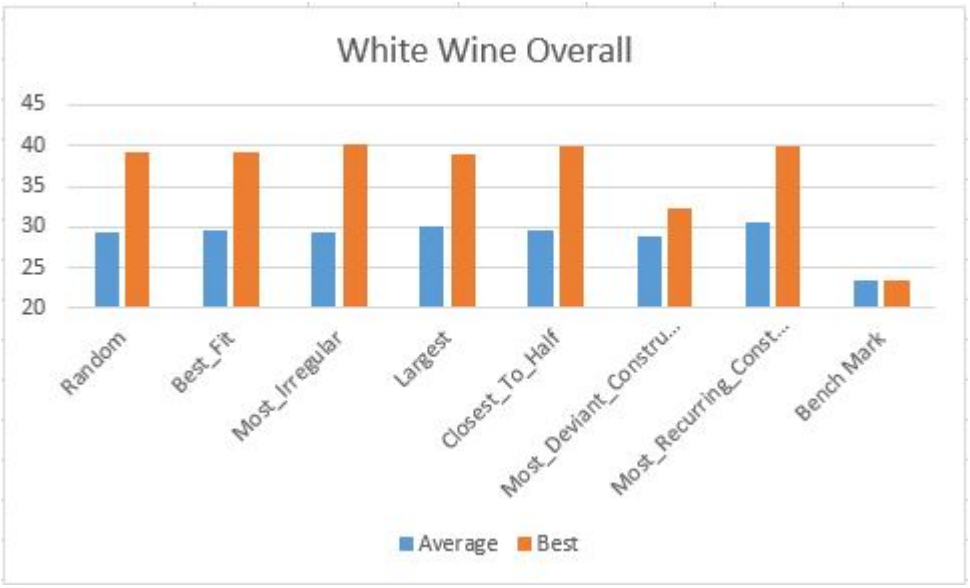

Figure 4.6: Graph of accuracy for the white wine dataset.

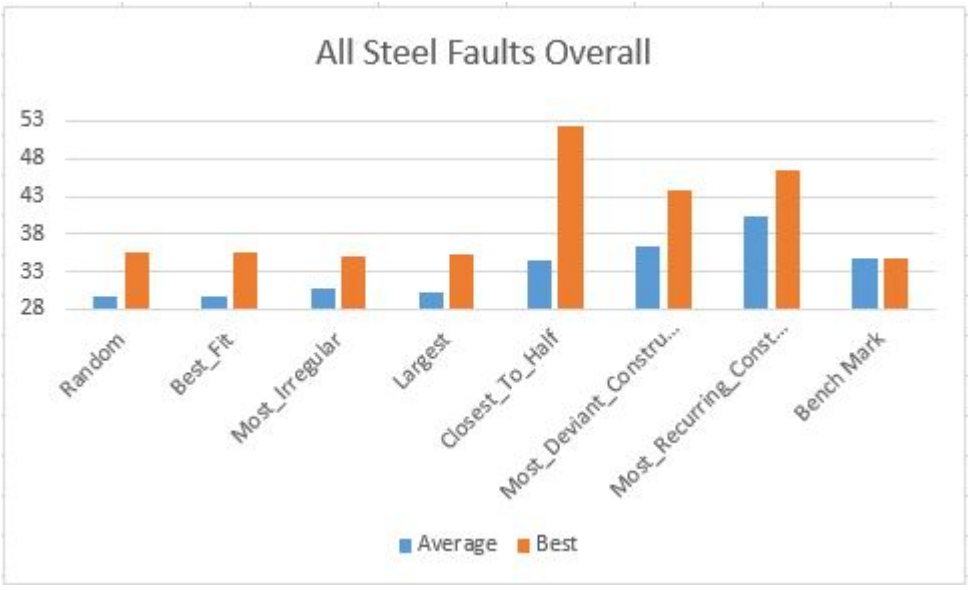

Figure 4.7: Graph of accuracy for the all steel faults dataset.

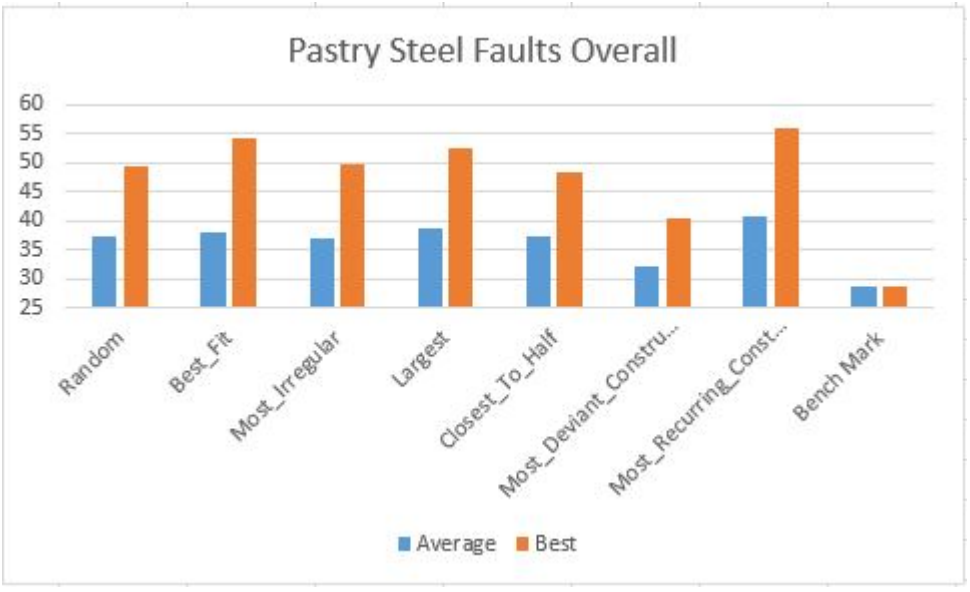

Figure 4.8: Graph of accuracy for the pastry steel faults dataset.

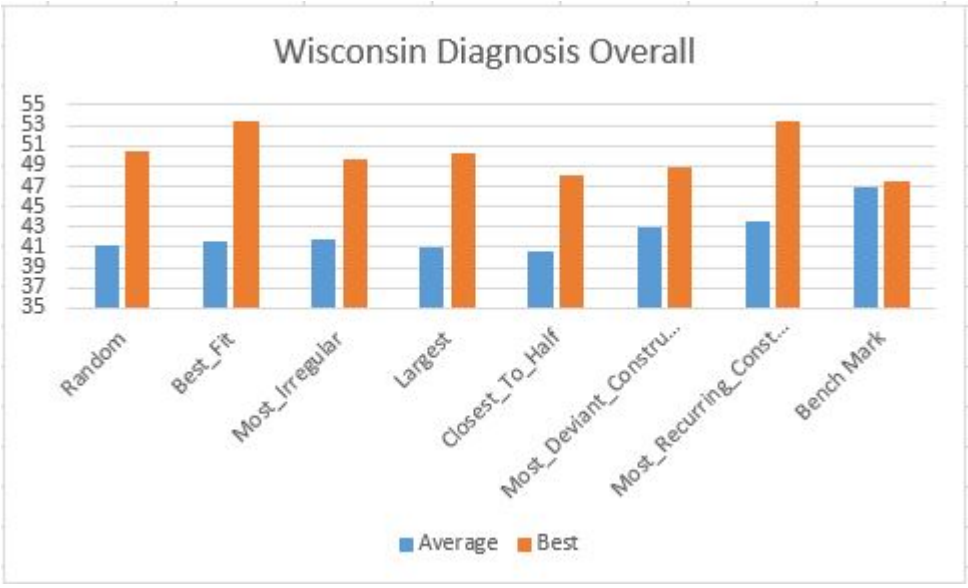

Figure 4.9: Graph of accuracy for the wisconsin diagnostic dataset.

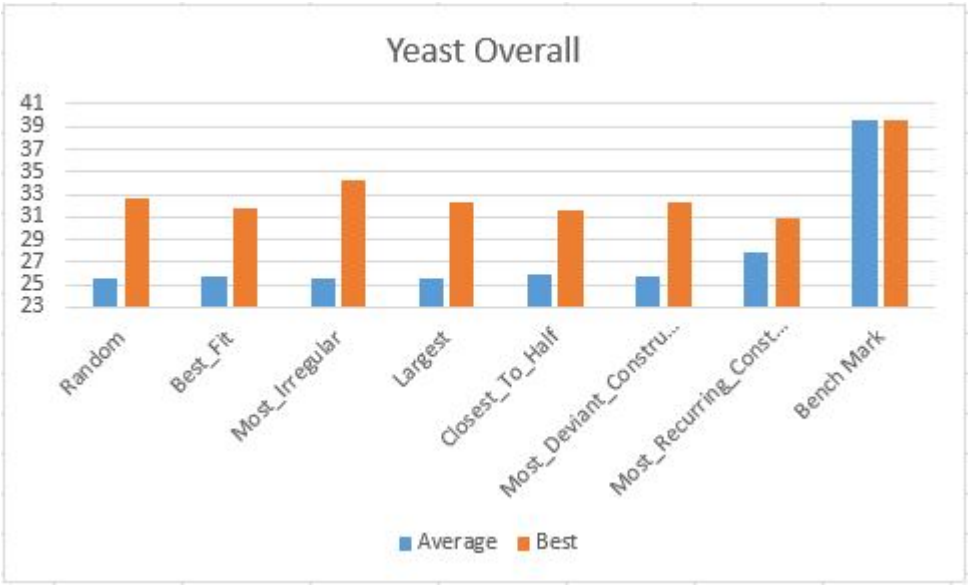

Figure 4.10: Graph of accuracy for the yeast dataset.

4.2 Methods that Perform Best

Examining Figures 4.1 - 4.10 certain patterns begin to emerge. The first of these patterns is that the selective methods (Random, Best Fit, Most Irregular, Largest, and Closest to half) perform quite similarly. The exception to this rule is the closest to half method, in some datasets (Steel Faults and Auto MPG) the closest to half method outperforms the other selective methods by a significant margin, both on average and in the best case.

The next pattern is an inverse relation between the success of the selective methods and the most deviant construction method. It seems that when the selective methods do on average worse than the bench mark the most deviant construction method outperforms the selective methods. Conversely when the selective methods outperform the benchmark the most deviant construction method does worse than the selective methods. In 8 of the 10 cases this relation is present, in the others (Dermatology, Wisconsin Diagnostic, and Yeast) the most deviant construction method performed very similarly to the selective methods.

Another pattern is the success of the most recurring construction method. For the average case the most recurring construction method had the best result for 8 of the 10 datasets. In the other two (Auto MPG and Dermatology) it was the second highest next to most deviant construction. For the best case the most recurring construction method had the best result in 6 of the 10 datasets. In 3 other 4 (Auto MPG, White Wine, Steel Faults) most recurring construction was the second or third best method. However in the Yeast dataset the most recurring construction method performed the worst, but only by a small margin.

A particularly observant reader may also notice the rather small difference between the average and best cases for the most deviant construction method. In 9 out of 10 cases this difference is least of any method, often by a large amount. This can be explained by the nature of the method. The most deviant construction method does not calculate witnesses of ε-irregularity, it instead constructs witness from those vertices that have degree differing from the average by the most. This means that the process of generating witnesses does not depend on ε at all, instead ε only determines our stopping condition. Since we still collected data for each value of ε, there were many data points which were very close together since, for example, the results for refinement factor 3 and $\varepsilon = 0.2$ are very close to refinement factor 3 and $\varepsilon = 0.6$.

4.3 Best Choice of Parameters

The large difference in the accuracy of the best case and average case for our seven Regularity Clustering methods highlights the importance of the choice of parameters (ε and refinement factor). In order to determine which values of ε and refinement factor produce above average results we examined the average accuracy of each method for each value of both parameters. Figures 4.11 - 4.24 are graphs depicting this analysis.

Let's first consider the choice of ε. Usually we consider ε-regular partitions to be of higher quality when a smaller ε is chosen. This is because as ε approaches zero the behavior of the partition approaches the expected value of a similar partition of a random graph. However it is not as obvious that smaller ε values will yield better results in Regularity Clustering (and many times they do not). This is because the algorithm will, in most cases, terminate before constructing an ε-regular partition. Since this is the case we would like for the algorithm to do as much work in each iteration as possible, in order to get as close to an ε-regular partition as possible. Witnesses of ε-irregularity for larger values of ε have larger lower bounds for both size and the difference in density and are thus presumably of higher quality. However, witnesses of large ε-irregularity are few in number and will occasionally not exist for certain pairs.

Since we have not developed a method of determining this value for any given dataset we have settled for analyzing the results of our data in an attempt to generalize our results to other datasets. The first observation of note is that the choice of ε makes very little difference for the Most Deviant Construction method. This makes sense because the most deviant construction method does not use ε when constructing its "witnesses". Instead, ε is used only when determining the stopping condition. The second observation of note is that $\varepsilon = 0.6$ yields very poor results for the random, best fit, most irregular, and largest methods but average results for the closest to half, most deviant construction and most recurring construction methods. The difference between the selective and constructive methods at the $\varepsilon = 0.6$ is likely explained by the difference in sizes of the chosen witnesses. The selective methods choose a witness of $\varepsilon = 0.6$ irregularity which must be at least 60% of the partition class while the constructive methods make "witnesses" of a much smaller size. We cannot explain the success of the closest to half method at the $\varepsilon = 0.6$ method and hypothesize that a few of our datasets happened to be shaped in such a way that this method worked well at this

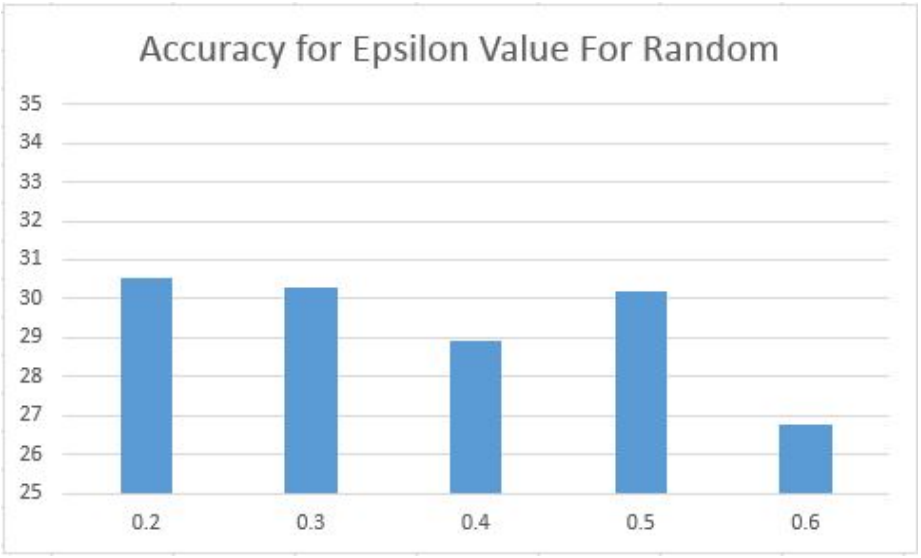

Figure 4.11: Accuracy of random method based on ε.

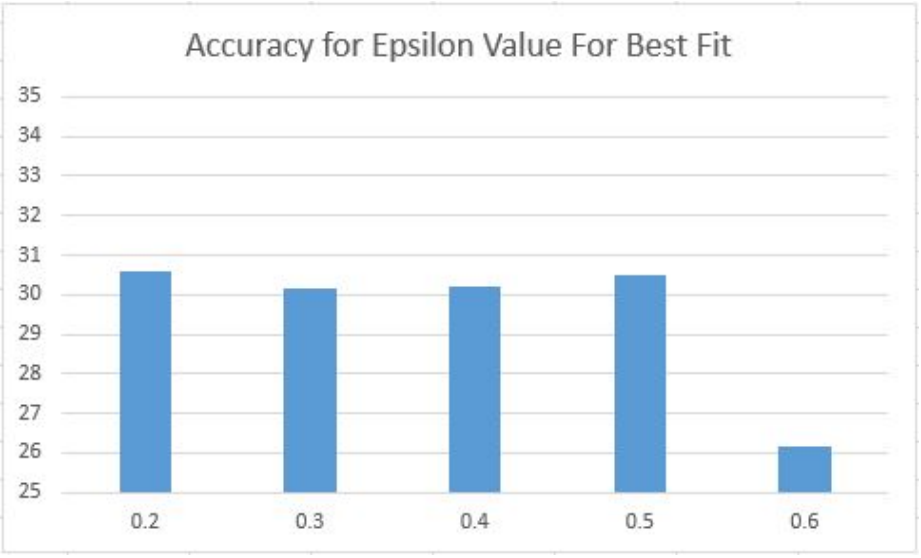

Figure 4.12: Accuracy of best fit method based on ε.

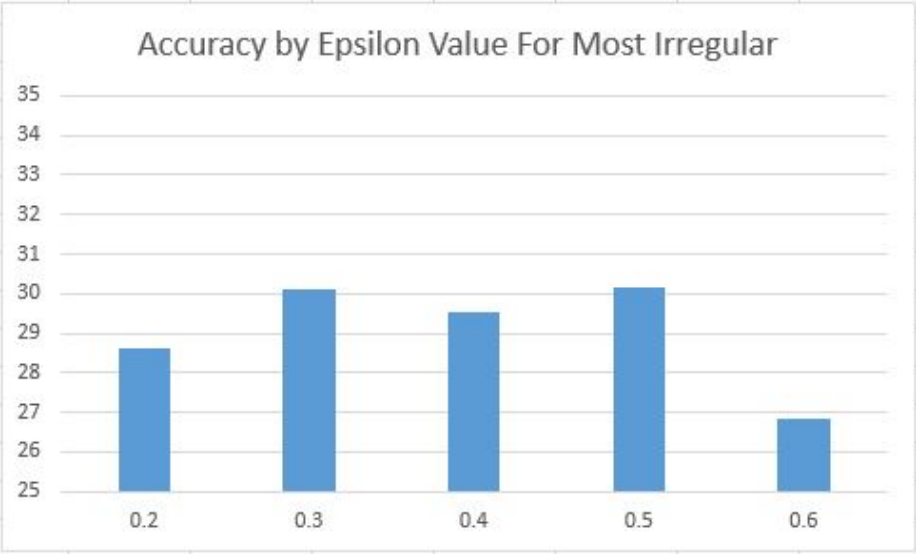

Figure 4.13: Accuracy of most irregular method based on ε.

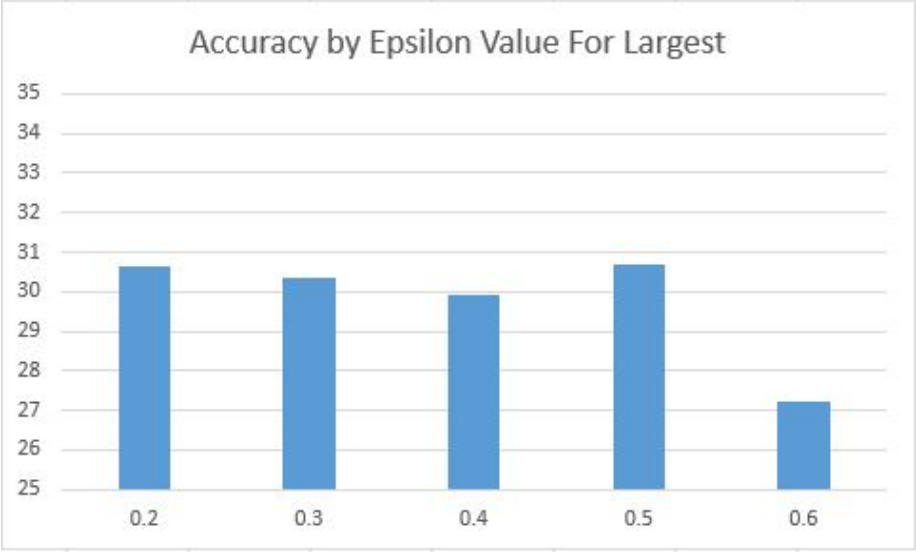

Figure 4.14: Accuracy of largest method based on ε.

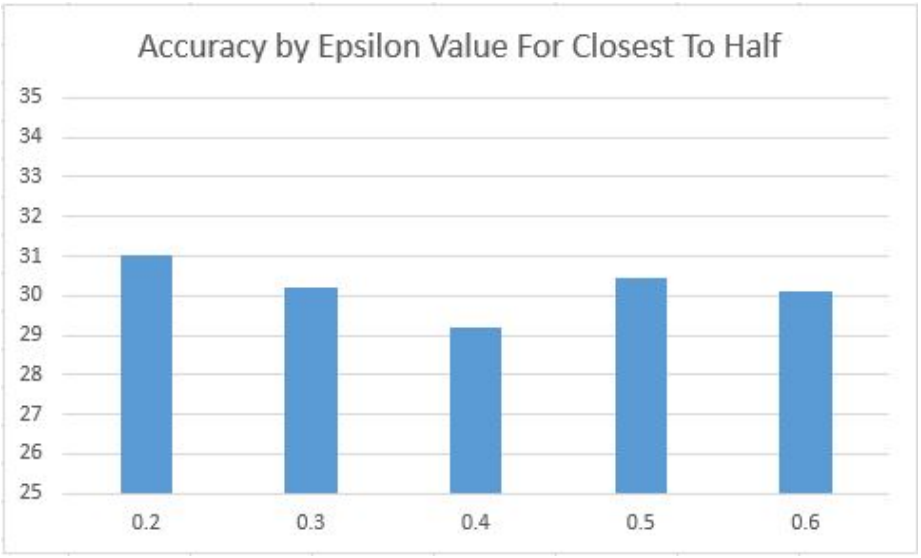

Figure 4.15: Accuracy of closest to half method based on ε.

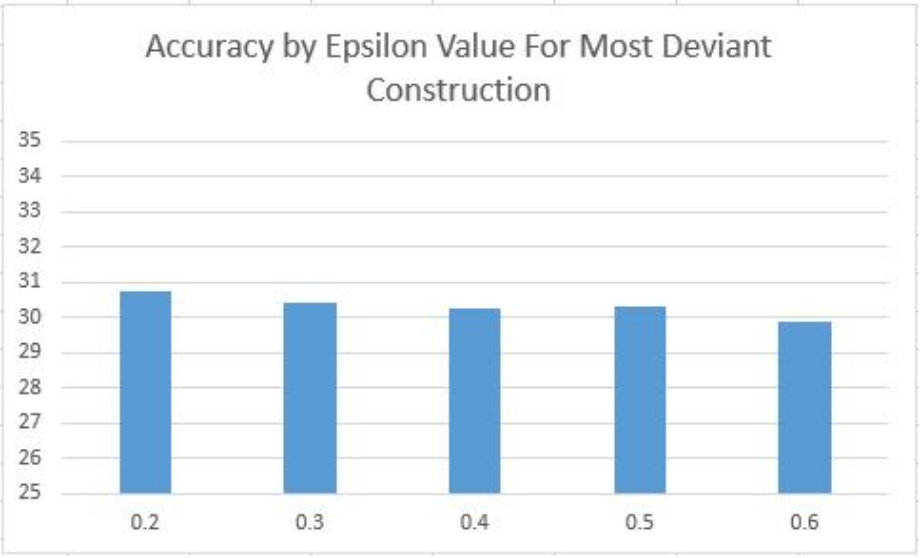

Figure 4.16: Accuracy of most deviant construction method based on ε.

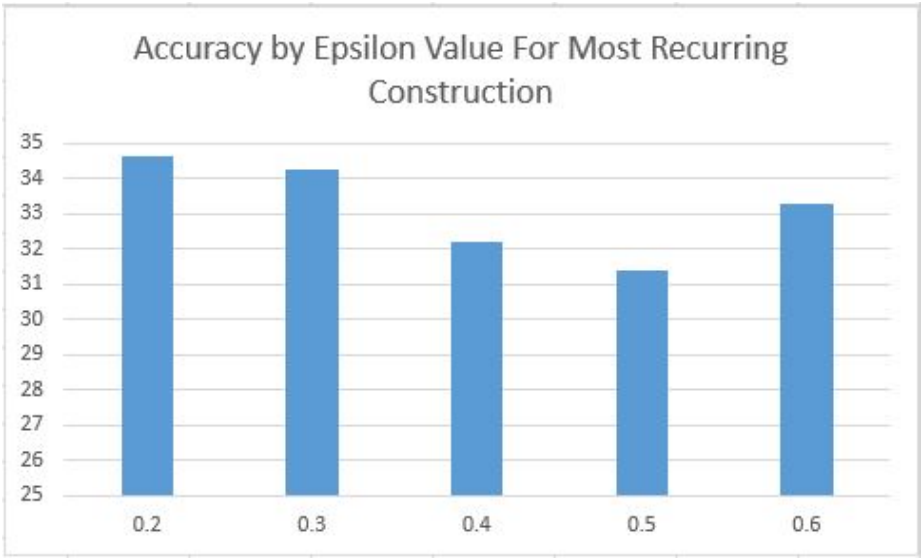

Figure 4.17: Accuracy of most recurring construction method based on ε.

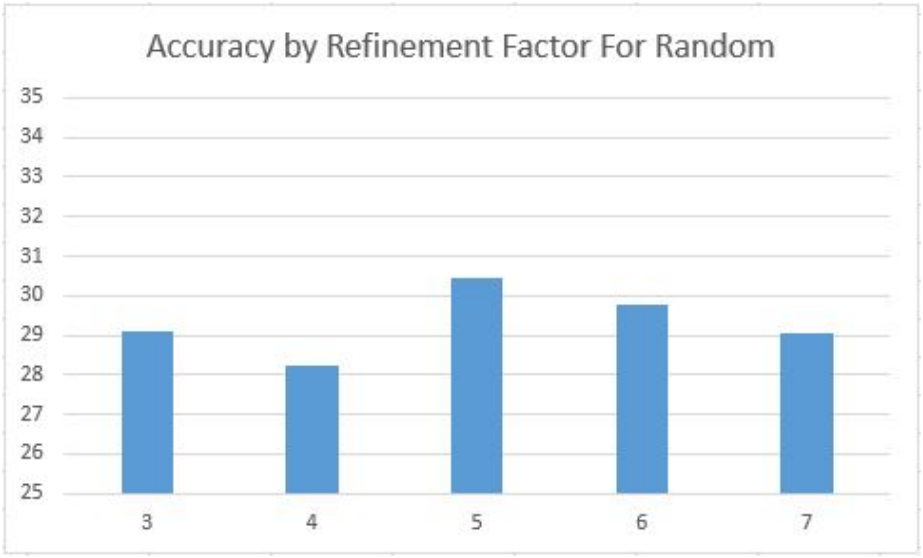

Figure 4.18: Accuracy of random method based on refinement factor.

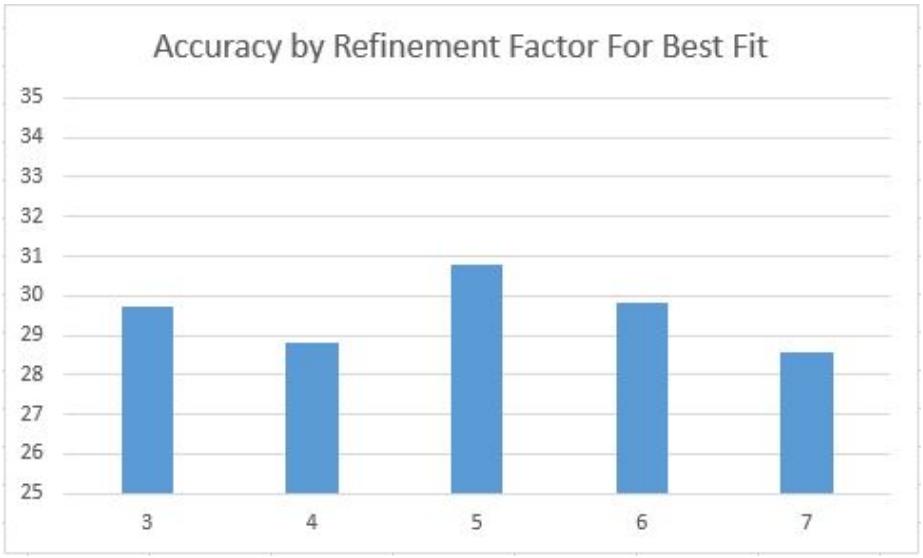

Figure 4.19: Accuracy of best fit method based on refinement factor.

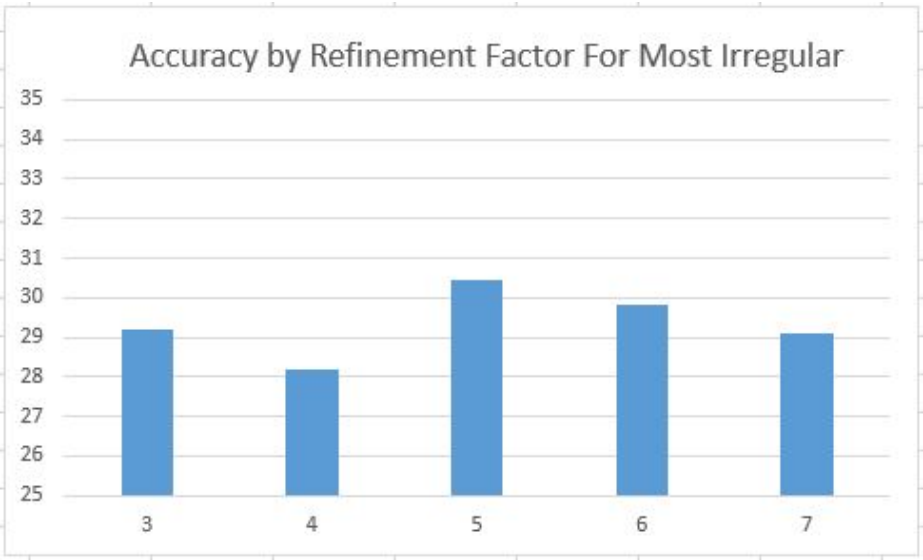

Figure 4.20: Accuracy of most irregular method based on refinement factor.

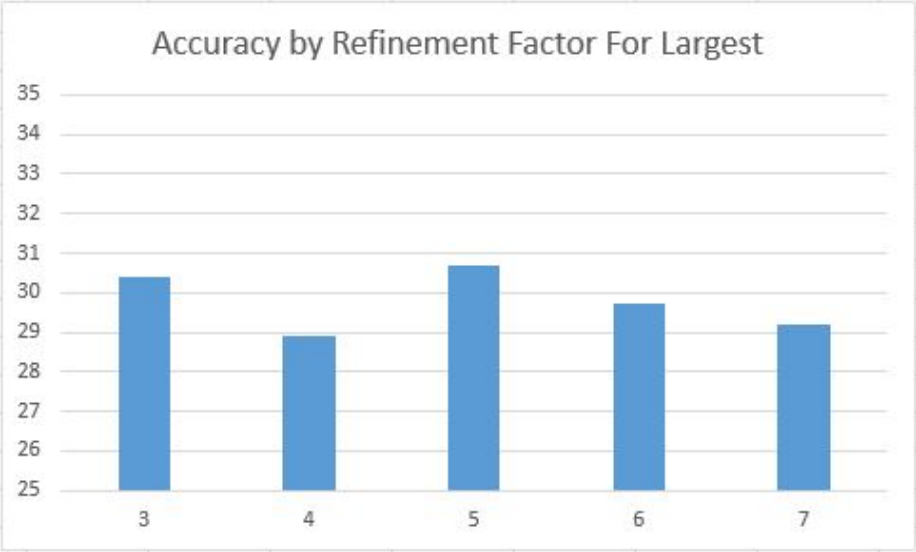

Figure 4.21: Accuracy of largest method based on refinement factor.

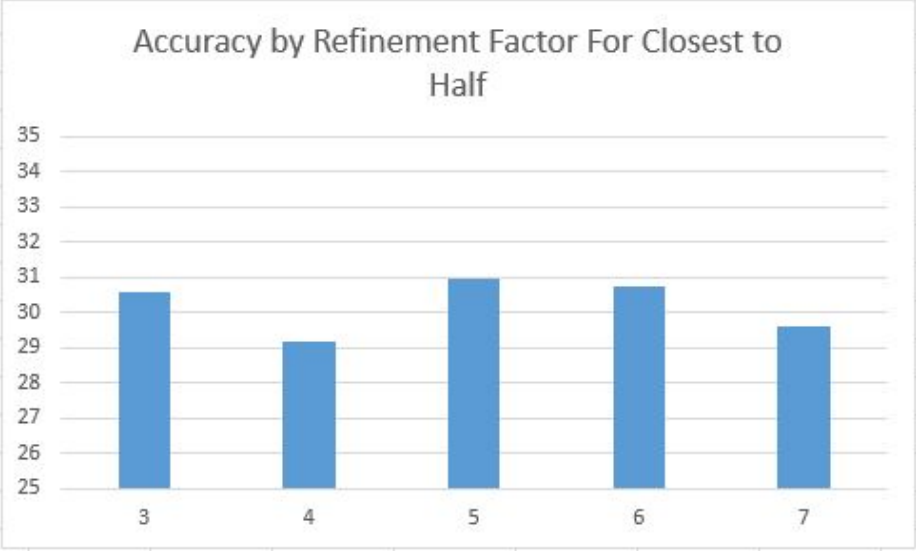

Figure 4.22: Accuracy of closest to half method based on refinement factor.

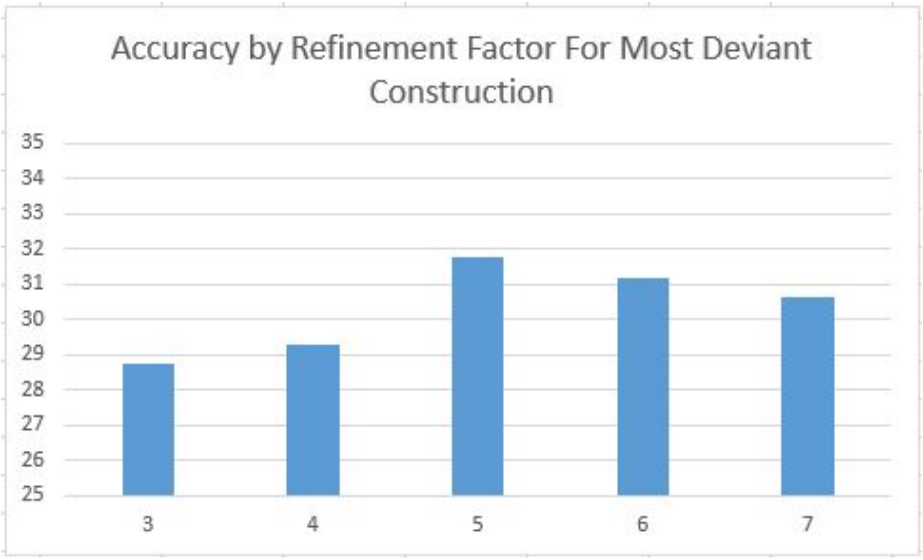

Figure 4.23: Accuracy of most deviant construction method based on refinement factor.

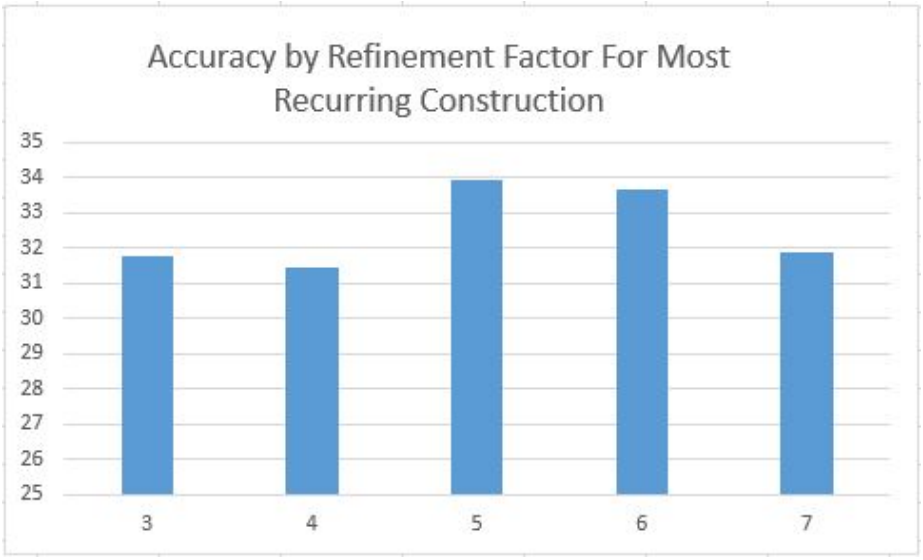

Figure 4.24: Accuracy of most recurring construction method based on refinement factor.

ε value.

Now lets consider the choice of refinement factor. By examining the graphs of the selective methods accuracy (Figures 4.18-4.22) we realize that they are very similar. This implies that the best choice of refinement factor is not at all dependent on which (selective) method you choose but rather on the dataset you are using. From these graphs we see that a refinement factor of 3 or 5 is often a good choice for selective methods. Similarly, we also see a pattern in the constructive methods. In both cases (most deviant and most recurring construction) refinement factor 3 and 4 produce poor results while refinement factor 5 produces the best result.

4.4 Conditions Under Which Regularity Clustering Perform Well

Figures 4.1-4.10 convey a wide variety of results. In some datasets (Haberman, Contraceptive Method Choice, Red Wine, White Wine, and Pastry Steel Faults) Regularity Clustering performs very well with every method outperforming the benchmark both on average and in best case scenarios. In others (Auto MPG, Steel Faults, and Wisconsin Diagnostic) Regularity Clustering performs worse on average but better for certain choices of ε and refinement factor. Still in others (Dermatology and Yeast) Regularity Clustering performs very poorly, where no matter the choice of ε and refinement factor we still perform worse than the benchmark. Since so little is known about the theoretical results of Regularity Clustering it would be very valuable to predict how well Regularity Clustering will perform on a particular dataset.

The similarities between the Red Wine and White Wine datasets and the Steel Faults and Steel Pastry Faults provide insight into possible factors. Regularity Clustering performed at 171% the benchmark for the White Wine dataset but 157% the benchmark for the Red Wine dataset. Since White Wine and Red Wine have the same attributes and the instances represent very similar things (red and white wine samples), this difference is likely caused by either the size of the data (White Wine has 4898 instances while Red Wine only has 1599) or by the fact that it is much harder to the quantify shape of the data (by this we could mean many things such as, how close the points are together and how evenly distributed the points are across the

target clusters).

Even more interesting is the difference in quality of our results on the Steel Faults dataset and the Steel Pastry Faults dataset. Regularity Clustering performed at 194% of the bench mark for the Steel Pastry Faults dataset but only 151% the bench mark for the Steel Faults datasets, even though its the same data! Recall that the difference between these sets. In Steel Faults we are trying to predict whether the steel fault is one of six specific types of faults or is some other kind of fault. In the Steel Pastry Faults dataset we try to predict if these same faults are pastry faults or some other type of fault. This seems to suggest that the number of target clusters plays a very important role in the success of Regularity Clustering.

To see if the rest of our data supported any of these theories we plotted the value $Max\{\frac{Method_i\_average}{Benchmark} : 1 \leq i \leq 7\}$ and $Max\{\frac{Method_i\_best}{Benchmark} : 1 \leq i \leq 7\}$ for each dataset versus an attribute of the dataset we wanted to test. The closer this plot is to linear with a non zero slope (or polynomial) the more our data suggests the attribute influences the quality of Regularity Clustering. We decided to examine combinations of four properties of the data: number of attributes, number of data points, number of target clusters, and distance from the expected distribution. By distance from the expected distribution we mean $\sqrt{\sum_{i=1}^{t} (ActualSize_t - ExpectedSize)^2}$, where t is the number of target clusters. We did not expect the graphs examining just the number of attributes, or number of target clusters to yield results as these parameters should also depend on the number of instances, however we included their graphs for completeness.

Figures 4.25 - 4.31 are the results of this analysis.

4.5 Hypotheses

In this section we will discuss the hypotheses we conclude from our data as well as a brief description of their theoretical merit.

4.5.1 Selection Methods

There are many different ways to go about selecting which witness(es) to use in the refinement process and the four we tested did slightly better than choosing a random witness. This seems to indicate that there is a best choice for a witness. Additionally, the closest to half method's occasional

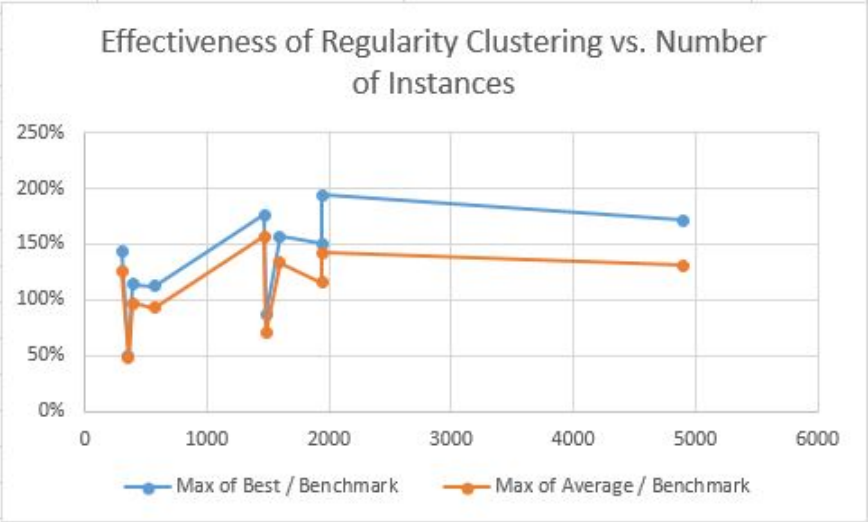

Figure 4.25: Comparing our results on each dataset to the benchmark based on the number of attributes.

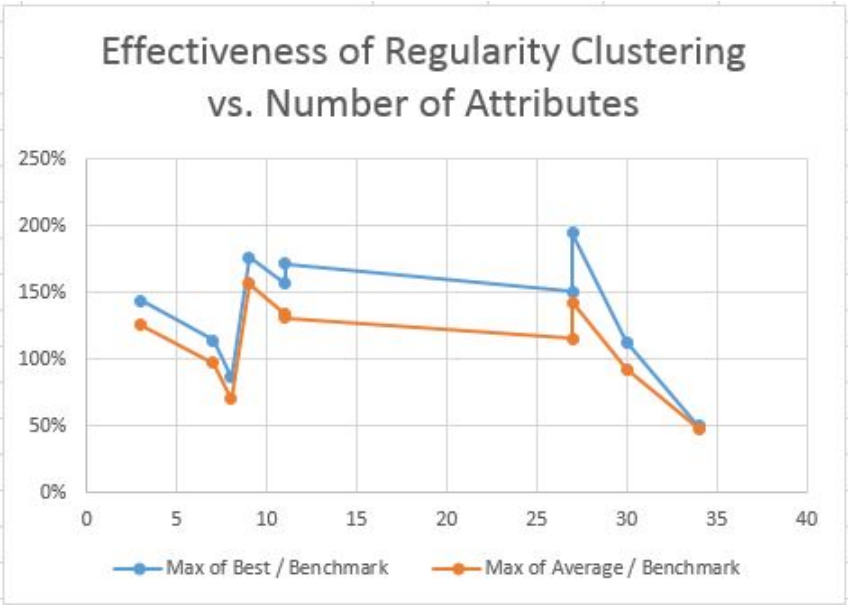

Figure 4.26: Comparing our results on each dataset to the benchmark based on the number of attributes.

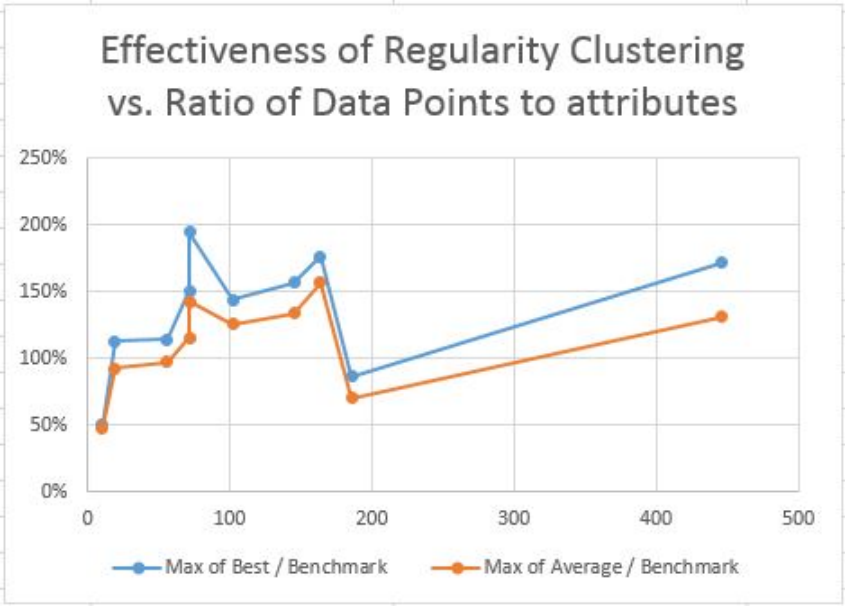

Figure 4.27: Comparing our results on each dataset to the benchmark based on the ratio of instances to number of attributes.

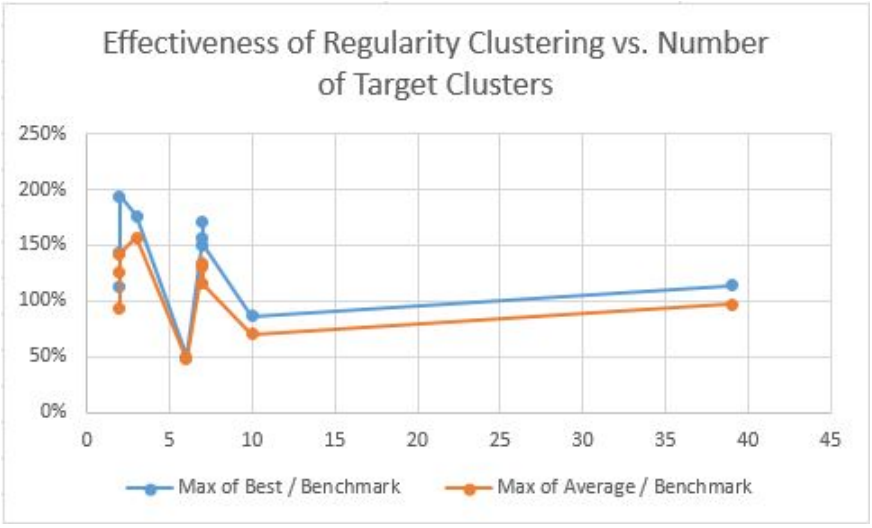

Figure 4.28: Comparing our results on each dataset to the benchmark based on the number of target clusters.

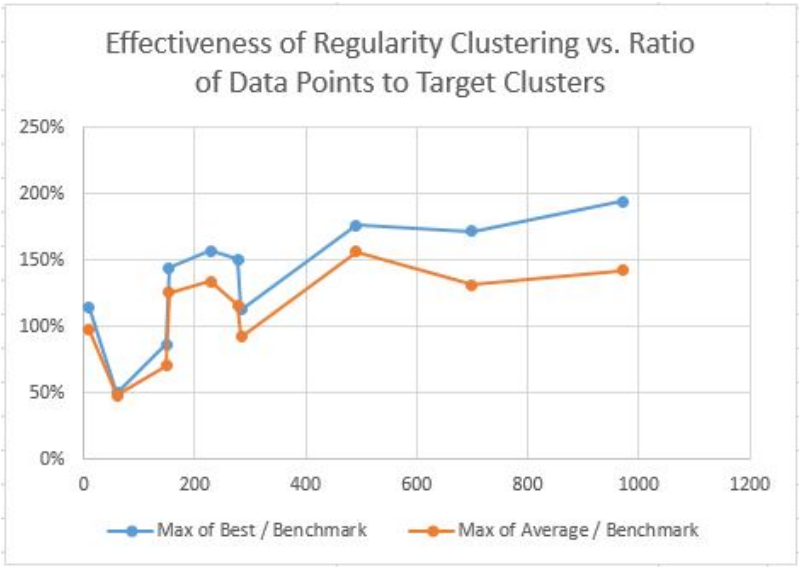

Figure 4.29: Comparing our results on each dataset to the benchmark based on the ratio of instances to the number of target clusters.

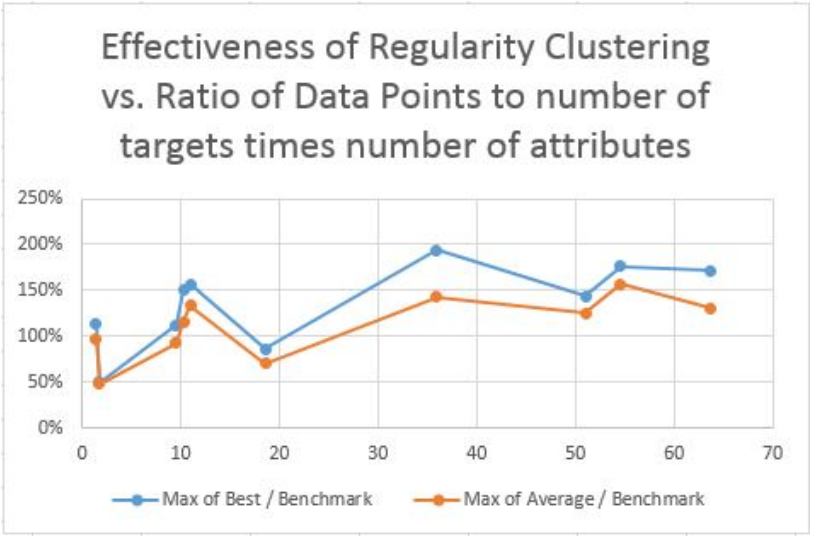

Figure 4.30: Comparing our results on each dataset to the benchmark based on the ratio of instances to target clusters times attributes.

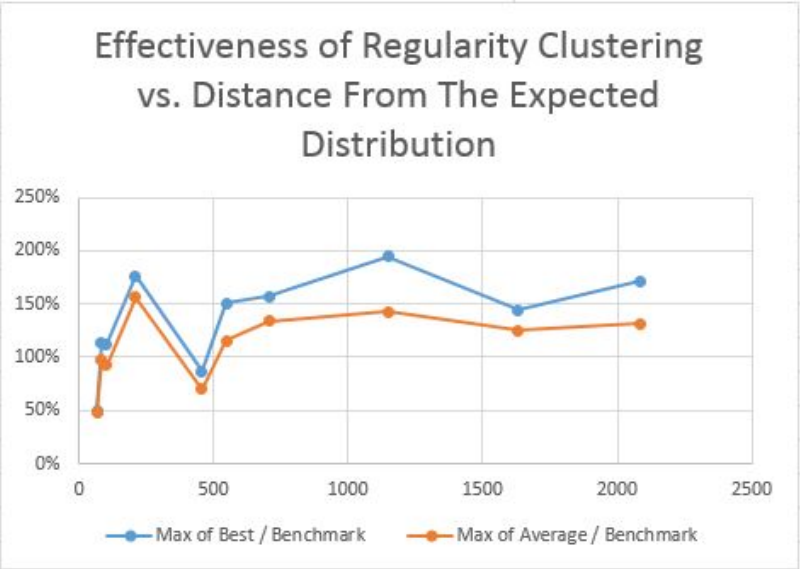

Figure 4.31: Comparing our results on each dataset to the benchmark based on the distance from the expected value.

huge success seems to indicate that separating a lot of vertices is important when attempting to choose the best witness. Beyond this it is difficult to say which of our methods is closest to choosing the best witness however our results show that each of these methods in general outperforms a random witness.

The existence of a best witness makes theoretical sense. While refining out each witness is guaranteed to increase the index by a certain amount there is nothing to say that a refining a witness will not increase the index by more than this amount. Additionally, witnesses have different properties such as size, how far the density of the witness differs from the average, and even how much the witness overlaps other witnesses. Our results show that these properties appear to make a difference in the quality of the witness.

4.5.2 Most Deviant Construction

It is our hypothesis that the most deviant construction method for generating witnesses will rarely, if ever be the best method choice. Our results seem to indicate that when the most deviant construction method outper-

forms the other methods for Regularity Clustering, Regularity Clustering is outperformed by the benchmark. Thus it seems that there is always a better choice than the most deviant construction method. On the other hand, it seems like a safe choice, in the sense that if you do not know if Regularity Clustering will perform well on your dataset you are not risking as much of the quality of your clusters. If it turns out Regularity Clustering does well on the data in question then you will still benefit from the use of the most deviant construction method over a technique other than Regularity Clustering. Conversely if Regularity Clustering does poorly on the data in question you will still do better than the other methods.

4.5.3 Most Recurring Construction

We hypothesize that of the Regularity Clustering methods we tested the most recurring construction method will in most cases produce the best results. Our data supports our claim, with the most recurring construction outperforming the other methods on average. We also hypothesize that a constructive method similar to most recurring construction will outperform most if not all selection methods. This is because selection methods depend upon the witnesses generated by algorithms which do not consider the quality of the witnesses they generate, merely that they are witnesses. Thus we predict that selecting even the best witness generated by the Alon et al. or Freize Kannan algorithms will not surpass the quality of the "witness" a constructive method could achieve.

4.5.4 Choice of Parameters

We realize that, with the exception of the method of choosing the most irregular witness (and largest by a very slight margin) $\varepsilon = 0.2$ yields the best results with $\varepsilon = 0.3$ very close behind. Figure 4.32 shows the average of our results across all datasets and all methods for each value of ε. From this graph and the results mentioned before we hypothesize 0.2 or 0.3 to be the best value for ε when nothing is known about the dataset in question.

We also hypothesize that the the choice of ε should depend heavily on the dataset in order to achieve the best results. Our idea to achieve the best (or at least very good) choice of ε would be to choose the largest value which yields at least one witness for each partition class. This would guarantee

that the witnesses used would be of the highest quality possible (assuming that witnesses of higher ε-irregularity are of higher quality).

On the other hand, the best choice of refinement factor appears to be independent of the method chosen. We hypothesize that the best choice of refinement factor is common among all selective methods. Additionally we have concluded that all of the methods we developed a refinement factor of 5 will generate the best results. Figure 4.34 shows the average of our results across all datasets and all methods for each value of refinement factor. This graph supports our claim.

4.5.5 When Regularity Clustering Performs Well

It is very hard for us to say anything with confidence in this regard as we only have ten datasets to work with. That being said our data appears to indicate that the size of the dataset, the number of attributes, and number of target clusters do not by themselves influence the results of Regularity Clustering. However, our data seems to roughly indicate that as the ratio of instances in the dataset to target clusters increases the quality of Regularity Clustering's results increase. It seems that a value of about 200 or higher indicates that Regularity Clustering is more likely to perform better than the bench mark.

This hypothesis may have some theoretical merit as well. We perform spectral clustering on the reduced graph that is produced as a result of the modified Regularity Lemma. A small number of instances compared to the number of target clusters will force the reduced graph to be too small to accurately spectral cluster into the target clusters. For instance if the reduced graph has less vertices than there are target clusters we are certainly in trouble. On the other hand, if this value is large the resulting partition will be fine enough for spectral clustering to be used effectively.

The ratio of instances in the dataset to the number of target clusters times the number of attributes also seems to roughly fit a positively sloped linear model. This might support our previous claims as including the number of attributes in the consideration seems to have reduced the correlation but not eliminated it. The correlation of the other attributes we tested to success are not supported by our data.

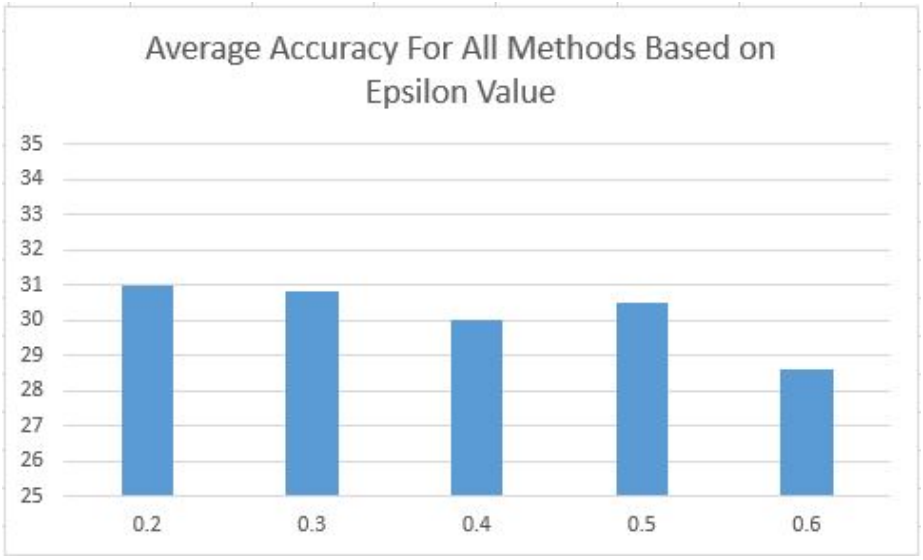

Figure 4.32: Accuracy of all methods based on ε.

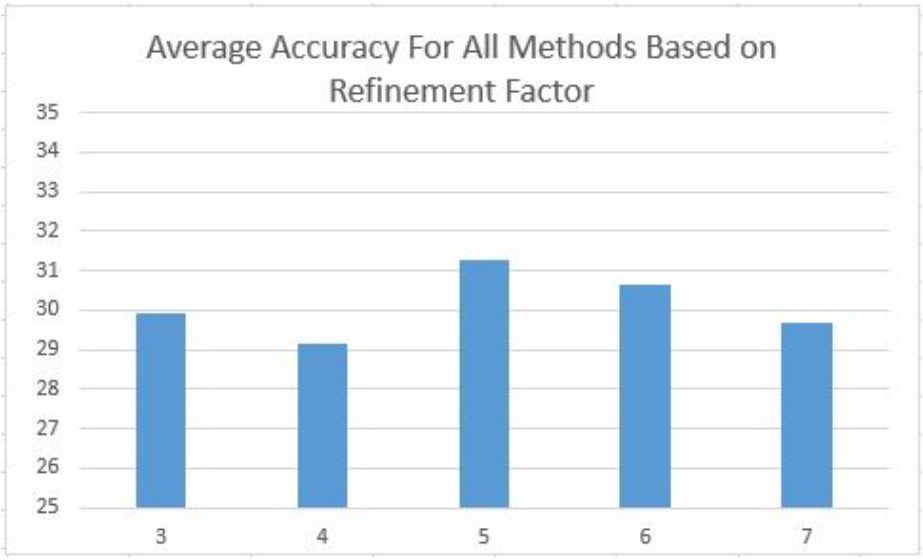

Figure 4.33: Accuracy of all methods based on the refinement factor.

Chapter 5

Conclusion

As data collection around the world increases at an exponential rate, the importance of data clustering continues to grow. Without a way to organize all of this data, the data becomes meaningless. Data clustering techniques give us a tool to predict anything from what online item a customer will most likely purchase to what stage of breast cancer a patient has. It has the power to teach us what factors are most important to a certain outcome and can lead to advances in nearly every field. Regularity Clustering has proved itself to be a very valuable tool and has a bright future among data clustering techniques.

The Regularity Lemma has been extremely influential in theoretical topics in mathematics and computer science which led to Szemerédi being awarded the Abel prize for this contribution. Although many believed Szemerédi's Regularity Lemma was purely theoretical, Regularity Clustering is a practical application of it. Our results show that the Regularity Lemma can be applied to the field of Big Data which suggests that its significance will continue to grow.

We discovered that the selection of which witness(es) is used during the refinement process had a large impact on the success of the clustering algorithm. Our selection methods performed better than when choosing a witness at random. This clearly shows that the selection methods provided a better witness for refinement.

We also determined that constructing a "witness" based on the witnesses produced in the Alon et al. algorithm can provide even better results. Our most recurring construction method outperformed all the other methods most of the time including standard spectral clustering. On average using most re-

curring construction significantly increased the accuracy of the final partition compared to other regularity clustering methods.

We concluded from our data that an ε value of 0.2 or 0.3 is best when nothing is known about the dataset being clustered. However, higher ε values can lead to even better results when there are enough witnesses created by the algorithm with that ε value.

The ideal choice of refinement factor varied considerably from one dataset to the next but remained consistent over different methods conducted on the same dataset. This clearly showed that choice of refinement factor depends mostly on the dataset rather than the method.

5.1 Future Work

There are many topics that came up during this project that would be interesting to pursue in the future. The stopping condition for the algorithm we used was determined by the original code we received from Sárközy, Song, Szemerédi, and Trivedi. The algorithm stopped once $\lceil \frac{class\_size}{2} \rceil \geq \frac{k_i}{\varepsilon}$ where k_i is the number of classes in the partition. Upon further investigation, it is unclear if this is an ideal stopping condition. With this stopping condition the number of iterations the algorithm performs is dependent on ε, which means that there exists a value such that ε's higher than that value cause the algorithm to perform an additional iteration. It is unclear what the correct number of iterations is for any given dataset but it is likely not the case that it should vary with epsilon. This is an area that we believe could be improved with further study.

We suggested in the analysis of our results that the best choice of ε may the largest value that ensures a witness for every partition class. We believe that modifying the algorithm to produce the "smartest" ε based on the data would be extremely effective in improvement the accuracy of the algorithm.

There were a few selection method we discussed but never followed through with implementing and testing. In one method we would choose the witness that overlaps the other witnesses most. This would be a witness that shares vertices with the largest amount of other witnesses. The benefit being that this set contains many vertices which lend witness to irregularity so by selecting this set we also select a maximal portion of the other witnesses. This is similar to the most recurring construction and has the benefit of being guaranteed to be an actual witness of irregularity, unlike the most recurring

construction method.

In another selection method we would find the largest set of witnesses which do not overlap to partition over. Recalling that the exponential refinement was caused by having to take the intersection of all the witnesses. If we are guaranteed the witnesses do not overlap then the refinement would be linear. This method will likely work better with higher refinement factors as there is more room for additional witnesses.

We also considered using witness pairs while refining which could lead to new results. Recall that an ε-irregular witness is generated by finding a subset of both ε-irregular sets such that the density of the subsets differs far from the density of the original set. The thought behind this choice of an ε-irregular witness is that the witnesses pair contains some of the information about why these vertices are irregular and should thus also be selected. However, when using the Alon et al. algorithm most pairs are a proper subset of the first set and the entirety of the second so it would not make any difference. Thus we expected this method would have a greater impact with the Frieze-Kannan algorithm.

In final summary, our work suggests methods for improving upon the already powerful Regularity Clustering technique. We also provide initial analysis on the conditions under which regularity clustering performs well. Our results suggest there is still unexplored potential in the field of Regularity Clustering and we present a number of avenues for future work exploring this potential.

Bibliography

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, R. Yuster, The Algorithmic Aspects of the Regularity Lemma. *Journal of Algorithms*, 16, (1994), pp. 80-109.

[2] C.L. Blake and C.J. Merz UCI repository of machine learning databases, 1998.

[3] Bache, K. and Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[4] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[5] B. Bollobás, P. Erdös, M. Simonovits, E. Szemerédi, Extremal graphs without large forbidden subgraphs, Annals of Discrete Mathematics 3 (1978), 29-41, North-Holland.

[6] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[7] R. Diestel, Graph Theory, 4th Electronic Edition 2010 Corrected reprint 2012 (2012) pp. 169-178.

[8] P. Erdös, A. Hajnal, V.T. Sós, E. Szemerédi, More results on Ramsey-Turán type problems, Combinatorica 3 (1983), 69-81.

[9] A. M. Frieze, R. Kannan, A simple algorithm for constructing Szemerédi's regularity partition. *Electron. J. Comb*, 6, (1999).

[10] W. T. Gowers. "A New Proof of Szemerédi's Theorem for Arithmetic Progressions of Length Four." Geometric And Functional Analysis 8.3 (1998): 529-51.

[11] W. T. Gowers. "The Work of Endre Szemerédi." Abelprize.no. The Abel Prize. Web. http://www.abelprize.no/c54147/binfil/download.php?tid=54060.

[12] Harold W. Kuhn, "The Hungarian Method for the assignment problem", Naval Research Logistics Quarterly, 2:8397, 1955.

[13] A. Ng, M. Jordan, Y. Weiss, On Spectral Clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), NIPS, MIT Press, 14, pp. 849-856, (2002).

[14] Norwegian Academy of Science and Letters, The, The Abel Prize. Online at http://www.abelprize.no/c54147/binfil/download.php?tid=54063 (2012).

[15] I. Z. Ruzsa, E. Szemerédi, Triple Systems with no six points carrying three triangles, Cominatorics (Keszthely, 1976), 18 (1978). Vol. II., 939-945. North-Holland, Amsterdam-New York.

[16] Gábor N. Sárközy, Fei Song, Endre Szemerédi, Shubhendu Trivedi, A Practical Regularity Partitioning Algorithm and its Applications in Clustering, arXiv:1209.6540, (2012).

[17] Semeion, Research Center of Sciences of Communication, Via Sersale 117, 00128, Rome, Italy. www.semeion.it

[18] E. Szemerédi, On sets of integers containing no four elements in arithmetic progression, Acta Math. Acad. Sci. Hung. (1969), 20: 89104

[19] E. Szemerédi, On graphs containing no complete subgraphs with 4 vertices (in Hungarian), Matematikai lapok 23 (1972), 111-116.

[20] E. Szemerédi, Regular Partitions of Graphs, Colloques Internationaux C.N.R.S. No 260 Problèmes Combinatoires et Théorie des Graphes, Orsay, pp. 399-401, (1976).