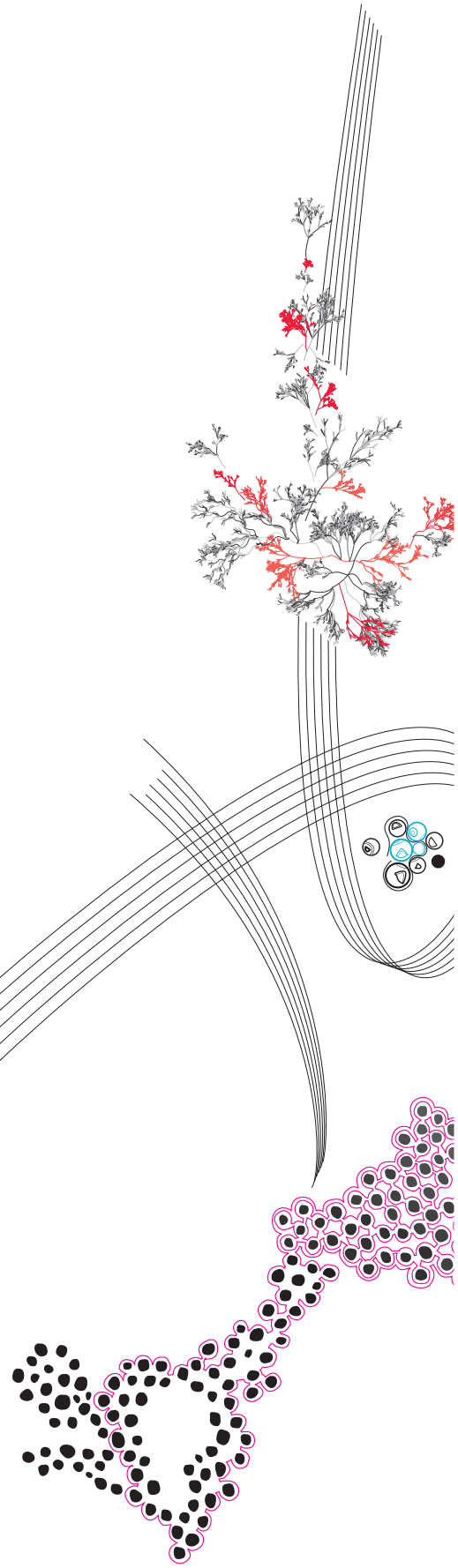

BSc Thesis Applied Mathematics

Clustering properties of random shortest path metrics based on edge expansion

D. M. van der Linden

Supervisor: B. Manthey

June, 2021

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Preface

I want to thank my supervisor Bodo Manthey for his guidance and I want to compliment him on his great eye for detail. I want to thank Lavinia Lanting and Bozhidar Petrov for the helpful talks and emotional support. And a special thanks to Bozhidar Petrov for doing the final proofread.

Clustering properties of random shortest path metrics based on edge expansion

D. M. van der Linden [Email: d.m.vanderlinden@student.utwente.nl]

June, 2021

Abstract

Probabilistic analysis for metric optimization problems has mostly been conducted on random Euclidean instances, but little is known about metric instances drawn from distributions other than the Euclidean. We consider the following model: take a connected graph where each edge has a weight that is independently drawn from an exponential distribution. The distance between two nodes is then the length of a shortest path (with respect to the weights drawn) that connects these nodes. Research on the clustering properties of complete graphs has been done by Bringmann et al. [1], and their findings have been generalized for non-complete graphs by Klootwijk et al. [4]. No research has been done however on the clustering properties of sparse graphs. This motivates our study of the clustering properties of expander graphs, which are sparse graphs a high connectivity. We have shown that if certain conditions apply, that the expected number of partitions required, for a clustering with clusters of at most diameter 4Δ, containing all vertices of a graph G, on n vertices, is $O(1 + n\,\mathrm{e}^{-\frac{1}{2}h(G)\Delta})$, where $h(G)$ is the edge expansion ratio.

Keywords: Expander Graph, Edge expansion, Graph, Clustering, Sparse, Random shortest paths, RSP, Random metrics, Expectation, Exponential Distribution.

1 Introduction

Simple heuristics often show a remarkable performance in practice for finding good solutions of optimization problems. Worst-case analysis often falls short of explaining this performance. Because of this, "beyond worst-case analysis" of algorithms has recently gained a lot of attention, including probabilistic analysis of algorithms [4].

The instances of many optimization problems are essentially a discrete metric space. Probabilistic analysis for such metric optimization problems has nevertheless mostly been conducted on instances drawn from Euclidean space, which provides a structure that is usually heavily exploited in the analysis. However, most instances from practice are not Euclidean. Little work has been done on metric instances drawn from other, more realistic, distributions [4]. Some initial results have been obtained by Bringmann et al. [1], who have used random shortest path metrics constructed using complete graphs to analyze heuristics. Some further results have been obtained by Klootwijk et al. [4], who have generalized those findings to non-complete graphs, especially Erdös-Rényi random graphs. In this paper we will expand on the work done by Bringmann et al. and by Klootwijk et al. by looking at the clustering properties of expander graphs and specifically edge expansion.

Consider a road network, where one can model the cities as vertices and the roads as edges. One could assign edge weights based on how long it takes to travel over a road, and then what would be relevant information is what the length of the shortest path is from one

node to another. This example might help to visualise and understand the more abstract content of this paper. We do not claim that the model in this paper is a good model for road networks in general. However we think that using expander graphs to model a road network is more realistic than using complete or almost complete graphs.

The traveling salesman problem (TSP) is a problem where one is tasked with finding a tour, or walk, that visits all nodes in a graph. The quality of the solution is judged by the length of the tour. The greedy heuristic is a heuristic solving this problem. This heuristic works by choosing its next vertex to visit to be the closest vertex (could one of several) that has not yet been visited. This heuristic shows remarkably good performance on the TSP in practice, while it's worst-case performance is quite poor.

The main result of this paper could be used to analyze the performance of this heuristic and might give insight as to why it performs as well as it does. We will use the remainder of the introduction to motivate how this can be done.

A clustering of graph $G = (V, E)$ is a partitioning of V into subsets, called clusters, such that every vertex of V is contained in exactly one cluster. The diameter of a cluster C_i is $\max_{u,v \in C_i} d(u, v)$, where $d(u, v)$ denotes the length of the shortest path between u and v.

Suppose we have a clustering $C = \{C_1, ..., C_{|C|}\}$ of a graph $G = (V, E)$ on n vertices, where each cluster is at most diameter Δ. Let $T = (v_1, ..., v_n)$ be the sequence of vertices that is obtained by applying the heuristic.

Let $i \in \mathbb{N}$ and $1 \leq i < n$. From the maximum diameter of each cluster it follows that, if there there exists a vertex v_m such that $m > i$, and v_i and v_m are in the same cluster, then the distance between consecutive vertices v_i and v_{i+1} is less then or equal to Δ. Suppose that the distance of the shortest path between vertex v_i and v_{i+1} is more than Δ. Then, there does not exist a vertex v_m where $m > i$, and v_i and v_m are in the same cluster. This can only occur once per cluster. Since there are $|C|$ clusters, the number of times the distance between two consecutive vertices of the sequence is more than Δ is bounded by $|C|$. This in turn can be used to obtain a bound on the performance of the greedy heuristic.

2 Notation and Model

We take over notation from paper by Klootwijk et al. [4]. We use $X \sim P$ to denote that a random variable X is distributed using a probability distribution P. $\text{Exp}(\lambda)$ is being used to denote the exponential distribution with parameter λ. In particular, we use $X \sim \sum_{i=1}^{n} \text{Exp}(\lambda_i)$ to denote that X is the sum of n independent exponentially distributed random variables having parameters $\lambda_1, ..., \lambda_n$. Let X be a random variable that is stochastically dominated by a random variable Y, i.e., we have $F_X(x) \geq F_Y(x)$ for all x, where F_X and F_Y are the cumulative distribution functions of X and Y, respectively. We denote this by $X \precsim Y$.

Let $\lfloor x \rfloor$ denote x rounded down to the largest integer below or equal to x.

2.1 Generalized random shortest path metrics

Given an undirected connected graph $G = (V, E)$ on n vertices, we construct the corresponding generalized random shortest path metric as follows. First, for each edge $e \in E$, we draw a random edge weight $w(e)$ independently from an exponential distribution with parameter 1. Second, we define the distances $d : V \times V \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ as follows: for every $u, v \in V$, $d(u, v)$ denotes the length of the shortest u, v-path with respect to the

drawn edge weights. If no such path exists we set $d(u,v) = \infty$, this does not occur in a connected graph. By doing so, the distance function d satisfies $d(v,v) = 0$ for all $v \in V$, $d(u,v) = d(v,u)$ for all $u, v \in V$ (symmetry), and $d(u,v) \leq d(u,s) + d(s,v)$ for all $u, s, v \in V$ (the triangle inequality). We call the distance function d obtained from this process a generalized random shortest path metric. Note that even though the graph G does not need to be a complete graph, the metric d is always complete in the sense that between each pair of vertices $u, v \in V$ the distance between u and v is defined $d(u,v)$.

2.2 Notation related to random shortest path metrics

We use the following notation within generalized random shortest path metrics: $\Delta_{\max} := \max_{u,v} d(u,v)$ denotes the diameter of the graph. Note that $\Delta_{\max} < \infty$ if and only if G is connected. $B_\Delta(v) := \{u \in V \mid d(u,v) \leq \Delta\}$ denotes the ball of radius Δ around v, i.e., the set containing all vertices at distance at most Δ from v. $\tau_k(v) := \min\{\Delta \mid |B_\Delta(v)| \geq k\}$ denotes the distance to the kth closest vertex from v (including v itself). Equivalently, one can also say that $\tau_k(v)$ is equal to the smallest Δ such that $B_\Delta(v)$ contains at least k vertices. Now, $B_{\tau_k(v)}(v)$ denotes the set of the k closest vertices to v. During our analysis, we will make use of the size of the cut induced by this set, which we will denote by $\chi_k(v) := |\delta(B_{\tau_k(v)}(v))|$, where $\delta(U) = \{uv \in E \mid u \in U \text{ and } v \in V \setminus U\}$ denotes the cut induced by U.

2.3 Expander related definitions

Loosely speaking, expander graphs are graphs with few edges which are highly connected. The formal definitions of expander graphs varies from paper to paper. We will use a definition based on edge expansion.

Definition 2.1 ([2], Definition 2.1). We define the edge expansion ratio $h(G)$ of graph $G = (V, E)$ on n vertices, to be

$$h(G) := \min_{\substack{0<|S|\leq \frac{n}{2} \\ S \subset V}} \frac{|\delta(S)|}{|S|}.$$

When considering connected graphs, the values of h are between 0 and roughly $\frac{n}{2}$. One end of the spectrum would be a where G is a tree on n vertices. Then, $h(G)$ is approximately equal to $\frac{2}{n}$. The other end of the spectrum is where G is a complete graph. If G is a complete graph on n vertices, then $h(G)$ is approximately equal to $\frac{n}{2}$.

A graph $G = (V, E)$ with n vertices is sparse if $|E| = O(n)$. Our definition of expander will be inspired by the definition by Hoory et al. ([3], Definition 2.2). There definition required that expander graphs are regular, we relaxed that condition and only require that expander graphs are sparse.

Definition 2.2. A sequence of sparse graphs $\{G_i\}_{i \in \mathbb{N}}$ of size increasing with i is a Family of Expander Graphs if there exists $c > 0$ such that $h(G_i) \geq c$ for all i. We then call the sequence c-expanding.

3 Results

Let $G = (V, E)$ be a graph on n vertices, defined as described in Section 2. From the definition of the edge expansion ratio we can obtain the following inequality.

Lemma 3.1. *If $k \in \mathbb{N}$, $v \in V$, and $0 < k \leq \frac{n}{2}$, then $k \cdot h(G) \leq \chi_k(v)$.*

Proof. Let $k \in \mathbb{N}$, $v \in V$, and $0 < k \leq \frac{n}{2}$. Then,

$$\begin{aligned} h(G) &= \min_{\substack{0<|S|\leq \frac{n}{2} \\ S\subset V}} \frac{|\delta(S)|}{|S|} && \text{(by definition)} \\ &\leq \frac{|\delta(B_{\tau_k(v)}(v))|}{|B_{\tau_k(v)}(v)|} && \text{(Restricting } S \text{ to be } B_{\tau_k(v)}(v)) \\ &= \frac{\chi_k(v)}{k}, \end{aligned}$$

where the last equality follows from the definition of $\chi_k(v)$, the definition of ball, and the definition of $\tau_k(v)$. Now for any given $k \in \mathbb{N}$ that satisfies $0 < k \leq \frac{n}{2}$, it follows that

$$h(G) \leq \frac{\chi_k(v)}{k}.$$

Since $k > 0$ this proves the statement. □

We can extend Lemma 3.1 to larger values of k. This is done in the following lemma. Note that the previous lemma is sufficient for proving the main result of this paper.

Lemma 3.2. *If $k \in \{1, ..., n-1\}$, $v \in V$, and $h(G)$ exists, then $\min(k, n-k) \cdot h(G) \leq \chi_k(v)$.*

Proof. Applying Lemma 3.1 proves the statement for $k \in \{1, ..., \lfloor \frac{n}{2} \rfloor\}$. Let $k \in \mathbb{N}$ satisfy $\frac{n}{2} \leq k < n$. Similar to the proof of Lemma 3.1, we have that

$$\begin{aligned} h(G) &= \min_{\substack{0<|S|\leq \frac{n}{2} \\ S\subset V}} \frac{|\delta(S)|}{|S|} \\ &= \min_{\substack{\frac{n}{2}\leq|S'|<n \\ S'\subset V}} \frac{|\delta(V \setminus S')|}{|V \setminus S'|} \\ &= \min_{\substack{\frac{n}{2}\leq|S'|<n \\ S'\subset V}} \frac{|\delta(S')|}{|V \setminus S'|} && \text{(Obtained from } |\delta(S')| = |\delta(V \setminus S')|) \\ &\leq \frac{|\delta(B_{\tau_k(v)}(v))|}{|V \setminus B_{\tau_k(v)}(v)|} \\ &= \frac{\chi_k(v)}{n-k}. \end{aligned}$$

Finally, since $(n-k) > 0$ it is implied that $(n-k) \cdot h(G) \leq \chi_k(v)$. □

Now that we have a lower bound on $\chi_k(v)$, we can prove the following about $\tau_k(v)$.

Lemma 3.3. *If $k \in \{1, ..., n-1\}$ and $h(G)$ exists, then*

$$\tau_k(v) \precsim \sum_{i=1}^{k-1} \text{Exp}(\min(i, n-i) \cdot h(G)).$$

Proof. From Lemma 3.2, the definition of stochastic dominance, and the fact that $\tau_k(v) \sim \sum_{i=1}^{k-1} \text{Exp}(\chi_i(v))$ (as proved by S. Klootwijk et al. [4]) the result follows. □

Furthermore, we can find a lower bound for the cumulative distribution

$$F_k(x) := \mathbb{P}(\tau_k(v) \leq x).$$

In order to prove this lower bound we need the following lemma about sums of exponentially distributed random variables.

Lemma 3.4 ([1], Lemma 3.2). *Let $X \sim \sum_{i=1}^{m} \mathrm{Exp}(c \cdot i)$, then for any $a \geq 0$ we have $\mathbb{P}(X \leq a) = (1 - \mathrm{e}^{-c \cdot a})^m$.*

Lemma 3.5. *If $k \in \{1, ..., \lfloor \frac{n}{2} \rfloor\}$, $h(G)$ exists, and $\Delta > 0$, then $F_k(\Delta) \geq (1 - \mathrm{e}^{-h(G)\Delta})^{k-1}$.*

Proof. By the definition of stochastic dominance, Lemma 3.3, and Lemma 3.4,

$$\begin{aligned} F_k(\Delta) &:= \mathbb{P}(\tau_k(v) \leq \Delta) \\ &\geq \mathbb{P}\left(\sum_{i=1}^{k-1} (\mathrm{Exp}(i) \cdot h(G)) \leq \Delta\right) \\ &= (1 - \mathrm{e}^{-h(G)\Delta})^{k-1}. \end{aligned}$$

Here the first inequality follows from Lemma 3.3 and the last equality follows from Lemma 3.4. □

In the proof of the main result of this paper we will discriminate between two types of vertices. Vertices that are close to a lot of other vertices, which are called 'dense' vertices, and vertices which are not, they are called 'sparse' vertices.

Definition 3.6. Let s be a value. We say $v \in V(G)$ is s-dense, if $|B_\Delta(v)| \geq s$, and s-sparse if it is not.

The following lemma yields a bound on the probability that a vertex v is k-sparse.

Lemma 3.7. *If $k \in \mathbb{N}$ satisfies $2 \leq k \leq \lfloor \frac{n}{2} \rfloor$ and $h(G), \Delta > 0$, then*

$$\mathbb{P}(|B_\Delta(v)| < k) < \frac{k-1}{\mathrm{e}^{h(G)\Delta}}.$$

Proof. Let $k \in \mathbb{N}$ satisfy $2 \leq k \leq \lfloor \frac{n}{2} \rfloor$ and $h(G), \Delta > 0$. Then,

$$\begin{aligned} \mathbb{P}(|B_\Delta(v)| \leq k) &= \mathbb{P}(\tau_k > \Delta) && \text{(by the definitions of } B_\Delta(v) \text{ and } \tau_k(v)) \\ &= 1 - F_k(\Delta) && \text{(by definition of } F_k(\Delta)) \\ &\leq 1 - (1 - \mathrm{e}^{-h(G)\Delta})^{k-1} && \text{(Lemma 3.4)} \\ &< \frac{k-1}{\mathrm{e}^{h(G)\Delta}} && \text{(Bernoulli's inequality).} \end{aligned}$$

□

Next we will prove the main result of this paper. We will show how to cluster the vertices of a graph, such that the expected number of clusters required is bounded, and the diameter of the clusters is bounded. A clustering of graph $G = (V, E)$ is a partitioning of V into subsets, called clusters, such that every vertex of V is contained in exactly one cluster. We will use Lemma 3.7 to prove the main result of this paper.

Theorem 3.8. *Let distance $\Delta > 0$, the number of vertices $n \geq 4$, and $\mathrm{e}^{h(G)\Delta} \geq 2$. Then, if we partition the vertices of G into clusters of at most diameter 4Δ, the expected number of clusters needed is $O(1 + n\,\mathrm{e}^{-\frac{1}{2}h(G)\Delta})$.*

Proof. Let $\Delta > 0$, and let $n \geq 4$, and $\mathrm{e}^{h(G)\Delta} \geq 2$. Let s_Δ be a value between 2 and $\lfloor \frac{n}{2} \rfloor$ that we will define more precisely later in the proof. For the remainder of this proof we shall refer to s_Δ-sparse and s_Δ-dense vertices as sparse and dense vertices, respectively.

Let us start by bounding the number of sparse vertices. The expected number of sparse vertices is $\sum_{v \in V} \mathbb{P}(|B_\Delta(v)| < s_\Delta)$. Since $n \geq 4$ and we have that $2 \leq s_\Delta \leq \lfloor \frac{n}{2} \rfloor$, we may apply Lemma 3.7, yielding that the expected number of sparse vertices is bounded from above by $n\frac{s_\Delta - 1}{\mathrm{e}^{h(G)\Delta}}$.

Now we will create clusters containing all dense vertices. Let H be an auxiliary graph, where the vertex set $V(H)$ of H, is equal to the set of all dense vertices of G. And the edge set $E(H)$, is defined by

$$uv \in E(H) \iff u, v \in V(H) \text{ and } d(u, v) \leq 2\Delta.$$

A subset S of $V(H)$ is an independent set of H if no two vertices of S are adjacent in G. An independent set S is a maximum independent of H, if H has no independent set S' with $|S'| > |S|$ [5].

Let $X = \{x_1, ..., x_{|X|}\}$ be a maximum independent set of H. Let $C'_1, ..., C'_{|X|}$ be our initial clusters where $C'_i = B_\Delta(x_i)$.

Now we obtain the final clusters $C_1, ..., C_{|X|}$. Firstly, let C_i contain all vertices of C'_i. Lastly, for each vertex v that is not in any of the clusters C'_i do the following. Let v be a vertex not in any of the clusters C'_i. We claim that there is a vertex x_i such that $d(v, x_i) \leq 2\Delta$. Suppose for the sake of contradiction that $d(v, x_i) > 2\Delta$ for all $x_i \in X$. If this were true then $X \cup \{v\}$ is an independent set of H. Since $|X \cup \{v\}| > |X|$ this contradicts that X is a maximum independent set of H, proving our claim that $d(v, x_i) \leq 2\Delta$. Let v be contained in one C_i such that $d(v, x_i) \leq 2\Delta$.

Since there are at most n dense vertices and each C_i contains at least s_Δ vertices there are no more then $\frac{n}{s_\Delta}$ clusters, with diameter at most 4Δ, required to contain all dense vertices.

Now we define $s_\Delta = \min(\lfloor \frac{n}{2} \rfloor, \lfloor \mathrm{e}^{\frac{1}{2}h(G)\Delta} \rfloor)$. Note that $2 \leq s_\Delta \leq \lfloor \frac{n}{2} \rfloor$. This leaves us with two cases.

Case 1 Suppose $\mathrm{e}^{\frac{1}{2}h(G)\Delta} \geq \lfloor \frac{n}{2} \rfloor$. Then $s_\Delta = \lfloor \frac{n}{2} \rfloor$. Therefore the expected number of sparse vertices is bounded from above by

$$\begin{aligned} n\frac{s_\Delta - 1}{\mathrm{e}^{h(G)\Delta}} &= \frac{n\lfloor \frac{n}{2} \rfloor - n}{\mathrm{e}^{\frac{1}{2}h(G)\Delta} \cdot \mathrm{e}^{\frac{1}{2}h(G)\Delta}} \\ &\leq \frac{n\lfloor \frac{n}{2} \rfloor - n}{\lfloor \frac{n}{2} \rfloor \lfloor \frac{n}{2} \rfloor} \\ &= O(1). \end{aligned}$$

The expected number of clusters, of at most diameter 4Δ, required to contain all dense vertices of G is bounded from above by $\frac{n}{s_\Delta} = \frac{n}{\lfloor \frac{n}{2} \rfloor}$ which is also is also $O(1)$.

Therefore the expected number of clusters, of at most diameter 4Δ, required to contain all vertices of G is $O(1)$.

Case 2 Suppose $\mathrm{e}^{\lfloor \frac{1}{2}h(G)\Delta \rfloor} < \lfloor \frac{n}{2} \rfloor$. Then $s_\Delta = \lfloor \mathrm{e}^{\frac{1}{2}h(G)\Delta} \rfloor$. Then the expected number of clusters, of at most diameter 4Δ, required to contain all vertices of G is bounded from above by

$$\frac{n}{s_\Delta} + n\frac{s_\Delta - 1}{\mathrm{e}^{h(G)\Delta}} = \frac{n}{\lfloor \mathrm{e}^{\frac{1}{2}h(G)\Delta} \rfloor} + n\frac{\lfloor \mathrm{e}^{\frac{1}{2}h(G)\Delta} \rfloor - 1}{\mathrm{e}^{h(G)\Delta}} = O(n\,\mathrm{e}^{-\frac{1}{2}h(G)\Delta}).$$

□

4 Concluding Remarks

We would like to start this section by giving some intuition as to why the main result is non-trivial. Let us suppose we have a c-expanding sequence of graphs, that satisfy what is described in Section 2. Further suppose we want to chose the maximum diameter of the clusters such that the expected number of clusters required for a clustering as described by Theorem 3.8 is $O(\sqrt{n})$. Now let $\Delta = \frac{\ln(n)}{c}$. Then, $\mathrm{e}^{\frac{1}{2}h(G)\Delta} = \mathrm{e}^{\frac{1}{2}\frac{h(G)}{c}\ln(n)} \geq \sqrt{n}$. For this choice of Δ and for n large enough Theorem 3.8 applies, yielding that the required number of clusters, of at most diameter $4\frac{\ln(n)}{c}$ is $O(\sqrt{n})$. Note this result does not hold for any sequence of sparse graphs. Take for instance a sequence of grid graphs.

To us the logical next step for researching this topic would be analyzing algorithms while making use of this clustering result. This could be done in a similar fashion to how this was done in the papers by Bringmann et al. [1] and by Klootwijk et al. [4]. The result proved by Klootwijk et al. [4] is as follows: If certain conditions apply and we partition the vertices into clusters, each of diameter at most 4Δ, then the expected number of clusters needed is bounded from above by $O(1 + n\,\mathrm{e}^{-\frac{1}{5}\alpha\Delta n})$, where α is a value between 0 and 1. The a major diffrence between the result of this paper and that one proved by Klootwijk et al. is that term e^{-n} in the order of the expected number of clusters required. Because of this difference the question remains, whether it is possible to prove similar results, when analysing algorithms applied to c-expanding sequences.

References

[1] K. Bringmann, C. Engels, B. Manthey, and B. V. R. Rao. Random Shortest Paths: Non-Euclidean Instances for Metric Optimization Problems. *Algorithmica*, 73:42–62, 2015.

[2] S. Hoory, N. Linial, and A. Wigderson. Expander Graphs and Their Applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[3] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[4] S. Klootwijk, B. Manthey, and S. K. Visser. Probabilistic analysis of optimization problems on generalized random shortest path metrics. *Theoretical Computer Science*, 874:106–111, 2021.

[5] WD Wallis. Graph theory with applications (ja bondy and usr murty), 1979.