CLUSTER ANALYSIS OF LOCAL CONVERGENT SEQUENCES OF STRUCTURES

JAROSLAV NEŠETŘIL AND PATRICE OSSONA DE MENDEZ

ABSTRACT. The cluster analysis of very large objects is an important problem, which spans several theoretical as well as applied branches of mathematics and computer science. Here we suggest a novel approach: under assumption of local convergence of a sequence of finite structures we derive an asymptotic clustering. This is achieved by a blend of analytic and geometric techniques, and particularly by a new interpretation of the authors’ representation theorem for limits of local convergent sequences, which serves as a guidance for the whole process. Our study may be seen as an effort to describe connectivity structure at the limit (without having a defined explicit limit structure) and to pull this connectivity structure back to the finite structures in the sequence in a continuous way.

CONTENTS

1. Introduction 2

Part 1. Preliminaries 7
2. Basic Definitions and Notations 7
3. Reduction from Local Formulas to Strongly Local Formulas 10

Part 2. Clustering Local Convergent Sequences 14
4. Negligible Sets and Sequences 14
5. Clusters and Pre-Clusters 17
  5.1. Clusters 18
  5.2. Universal Clusters 24
  5.3. Pre-Clusters 25
  5.4. Expanding Clusters 26
6. Clustering and the Cluster Comb Lemma 30
7. The Clustering Problem 36

Date: October 14, 2016.

2010 Mathematics Subject Classification. Primary 03C13 (Finite structures), 03C98 (Applications of model theory), 05C99 (Graph theory), 06E15 (Stone spaces and related structures), Secondary 28C05 (Integration theory via linear functionals).

Key words and phrases. Graph and Relational structure and Graph limits and Structural limits and Radon measures and Stone space and Model theory and First-order logic and Measurable graph.

Supported by grant ERCCZ LL-1201 and CE-ITI, and by the European Associated Laboratory “Structures in Combinatorics” (LEA STRUCO) P202/12/G061.
1. INTRODUCTION

Cluster analysis (being an established part of statistics, computer science and mathematics) is a core method for database mining. It initiated in the thirties in social sciences, particularly in anthropology and psychology. While the abstract notion of a cluster is somehow vague, some canonical types of cluster models have been considered, which allow to construct meaningful partitions of large data sets. Among these models, let us mention two principal extreme models: density models — where clusters correspond to connected dense regions, and distribution models — where clusters are defined by means of statistical distributions. For a comprehensive review of cluster analysis, we refer the reader to [9].

In this paper — which extends and precise some ideas introduced by the authors in [20] to study structural limits of trees — we propose a novel approach based on an interplay of these two models: knowing a limit statistical distribution associated to structures in a convergent sequence, we compute the parameters driving a density clustering of each of the structures in the sequence, in a seemingly “continuous” way. We believe that the cluster analysis presented here has a broader impact than the analysis of structural limits (which was our original motivation), and that it highlights a duality of the density and distribution models. Our analysis found immediate applications to the study of structural limits and we hope that more will come.

The convergence notion we use is the convergence of the distribution of the local properties of random vectors of elements. The limit distribution is used to drive a segmentation process, which can be seen as a marking of the elements of each structure in the sequence. The consistency of these markings is ensured by the requirement that the sequence of marked structures is still local convergent (see Fig 1 for a schematic visualization of this segmentation method).

Our approach is a natural one: if instead of considering a single snapshot of an evolving system we consider a significant part of the full movie, then clusters appear in a more obvious way, and meaningful parameters are much more easily defined and estimated. However the details are involved and lead to a new taxonomy.

Note that the notion of convergence considered here is a generalization of the notion of local convergence introduced by Benjamini and Schramm for graphs with bounded degrees [3]. In the general structural setting, introduced by the authors in [19], there
is no restriction on the degrees of the considered graphs or structures. (We refer the reader to [17] for an in-depth study of the classical notions of graph limits.) Informally, a sequence $\mathbf{A} = (\mathbf{A}_n)_{n \in \mathbb{N}}$ of structures is \textit{local convergent} if the probability $\langle \phi, \mathbf{A}_n \rangle$ (the Stone pairing of $\phi$ and $\mathbf{A}_n$) of satisfaction of every local first-order formula $\phi$ in structure $\mathbf{A}_n$ (for a random assignment of the free variables) converges as $n$ grows to infinity. (Recall that a \textit{local formula} is a formula whose satisfaction only depends on a bounded neighborhood of its free variables.) The limit of a local convergent sequence can thus be described by the (infinite) vectors of limit satisfaction probabilities $\lim_{n \to \infty} \langle \phi, \mathbf{A}_n \rangle$ indexed by all local first-order formulas $\phi$. This can also be represented as a probability measure, as stated by the general representation theorem (Theorem 3), in a way extending Aldous-Hoover representation of left limits of dense graphs by infinite exchangeable graphs [1, 13] and Benjamini-Schramm representation of local limits of graphs with bounded degree by an unimodular distribution on rooted connected countable graphs [3].

Our cluster analysis allows to meaningfully partition the structures in a local convergent sequence into dense connected clusters (plus an additional residual sparse cluster). It also show how this clustering is related to an imaginary connectivity structure of every limit structure one may define. Structural limits define convergence by means of statistics of first-order formula satisfaction. For various fragments of first-order logic we obtain various types of convergence. (In fact standard notions of left convergence [6, 18] and local convergence of graphs with bounded degrees [3] fall into this scheme.) It follows that limit objects one can construct share and express many interesting first-order properties. Can one say something about second order properties? Among them the most prominent is connectivity. What can we say about the asymptotic component structure of graphs in a convergent sequence? Note that the existence of a limit object does not help too much, as connectivity properties are not \textit{a priori} continuous.

In this paper we provide an analysis of the connectivity structures of each structure in a convergent sequence, leading to a connectivity analysis of every limit object one
can consider (ultraproduct endowed with Loeb measure [19], modeling [14], graphing, etc.) hence justifying a posteriori the relevance of satisfaction statistics for each connected component of a limit object with non-zero measure, and to the possibility to extract in the structures of the sequence the part converging to this particular connected component. Our assumptions are minimal: the analysis only uses the convergent sequence and the (unique) limit distribution associated to it by the representation theorem proved in [19].

Let us take time for a more detailed description both of our main result (Theorem 1) and of the main difficulties that we have to overcome to prove it. The first (surprising, at least at first glance) aspect, which already appears when considering Benjamini-Schramm limit of connected graphs with bounded degrees, is that the limit of a sequence of connected graphs needs not to be connected: if $\mathcal{G} = (\mathcal{G}_n)_{n \in \mathbb{N}}$ is a local convergent sequence of finite connected graphs with degree at most $D$ and with orders growing to infinity, then for every integers $k, r$ the probability that a random subset of $k$ vertices contains two vertices at distance at most $r$ tends to 0, which ultimately shows that the limit cannot have finitely many connected components. Actually every limit graphing will have uncountably many connected components.

When considering general local convergent sequences of finite structures, even if we don’t have an explicit limit structure, it makes sense to talk about the limit connected components and some of their properties. For instance, we prove that it is possible to determine the measure of all the limit connected components and, for
those with non-zero measures, their associated statistics. This is basically done by using Fourier analysis. Using this information, we prove that it is possible to track the limit connected components back to the structures of a local convergent sequence, by marking consistently the elements of all the structures in the sequence (see also schematic Fig. 2). The component structure of the limit is very complex and it has been repeatedly asked as a problem (by Lovász and others) how sequences of connected structures disconnect at the limit. Here we solve this problem at a general level, by showing that we can trace limit connected components with positive measure back in the sequence and how they gradually disconnect themselves from the remaining of the structures.

This analysis leads to interesting new notions: globular cluster (corresponding to a limit non-zero measure connected component), residual cluster (corresponding to all the zero-measure connected components taken as a whole), and negligible cluster (corresponding to the stretched part connecting the other clusters, which eventually disappears at the limit). Fig 3 summarizes schematically the interplay of the newly introduced notions. The marking of each of all these types of clusters will be explained in the second part of the paper. But let us mention that the main issue here is that we require that the marking of all these (countably many) clusters should preserve local convergence. This means that even if we consider local formulas using these marks, the satisfaction probabilities will still converge.

The main result of this paper reads as follows:
Theorem 1. Let $A$ be a local convergent sequence of $\sigma$-structures. Then there exists a signature $\sigma^+$ obtained from $\sigma$ by the addition of countably many unary symbols $M_R$ and $M_{i,j}$ ($i \in \mathbb{N}$, $1 \leq j \leq N_i$) and a clustering $A^+$ of $A$ with the following properties:

- For every $i \in \mathbb{N}$, $(\bigcup_{j=1}^{N_i} M_{i,j}(A_n^+))_{n \in \mathbb{N}}$ is a universal cluster;
- For every $i \in \mathbb{N}$ and every $1 \leq j \leq N_i$, $(M_{i,j}(A_n^+))_{n \in \mathbb{N}}$ is a globular cluster;
- Two clusters $(M_{i,j}(A_n^+))_{n \in \mathbb{N}}$ and $(M_{i',j'}(A_n^+))_{n \in \mathbb{N}}$ are interweaving if and only if $i = i'$;
- $(M_R(A_n^+))_{n \in \mathbb{N}}$ is a residual cluster.

(all undefined notions are explained below.)

The paper is organized in three parts, each subdivided into sections:

In the first part of the paper we introduce (in Section 2) the main definitions and notations used in this paper, and present (in Section 3) a reduction argument showing that considering strongly local formulas (that is local formulas whose satisfaction requires that all the free variables are assigned to vertices which are close) is sufficient to compute (exact) statistics component-wise, possibly after deletion of some set with negligible impact. For this purpose, a “weak algebra” of strongly local formulas is developed.

The second part of the paper is devoted to the theoretical study of an abstract notion of “cluster”. Section 4 is devoted to the study of sets with negligible impact, called negligible sets and to sequence of more and more negligible sets, called negligible sequences. Deletion of subsets forming a negligible sequence does not change the limit statistics of a local convergent sequence. Ultimately, our goal is thus to consider a negligible sequence, whose deletion will disconnect the graphs in the sequence into clusters. The formal notion of a cluster is discussed in Section 5. For us, a cluster will be a “continuous” sequence of subsets that correspond to a “stable entity” and that is “well separated” from the rest of the structure. This is expressed by the property that marking a cluster (formalized by considering a lift) preserves the local convergence, and that the frontier of a cluster forms a negligible sequence. Several types of clusters are defined and discussed in this section, in particular universal clusters and strongly atomic clusters. These last clusters corresponding to expanding parts of the structures in a local convergent sequence, and their properties are close to those of expander graphs. It follows from the definition of a cluster that iteratively marking finitely many clusters preserves local convergence. However, if we want to mark countably many clusters then the situation becomes more tricky. The conditions under which countably many clusters can be marked, sometimes modulo a limited modification, is discussed in Section 6, and is the purpose of the Cluster Comb Lemma (Lemma 26).

Particular clusters are intrinsically defined by the local convergence, which allow to mark dense spots in the structures of a local convergent sequence. These clusters,
called globular clusters, ultimately represent the non-zero measure imaginary connected components of the limit. To the opposite, a residual cluster represents a group of zero-measure imaginary connected components.

The third part of the paper is devoted to effective density clustering into countably many clusters and a residual cluster. In Section 8 we review the general representation theorem for limits of convergent sequences, and prove a general random rooting theorem using Fourier analysis. This result allows us to compute the spectrum of the sequence, from which we derive the asymptotic measures of the globular clusters. Using these informations, the actual computation of the clustering is done, and we deduce a complete characterization of all the globular clusters of the sequence, the computed globular clusters serving as a “globular basis”. in Section 9.

Part 1. Preliminaries

2. Basic Definitions and Notations

The theory of graph (and structure) convergence gained recently a substantial attention. Various notions of convergence were proposed, adapted to different contexts. Let us mention:

- the theory of dense graph limits [7, 18] based on the notion of left convergence,

These approaches have been (partly) unified by the authors in the setting of structural limits [19]. This last approach relies on a balance of model theoretic and functional analysis aspects. For a signature $\sigma$ and a fragment $X$ of the set of first-order formulas over the language generated by $\sigma$, we define for a finite $\sigma$-structure $A$ and a formula $\phi \in X$ with free variables $x_1, \ldots, x_p$ the Stone pairing of $\phi$ and $A$ as

$$\langle \phi, A \rangle = \frac{|\phi(A)|}{|A|^p},$$

where $\phi(A) = \{(v_1, \ldots, v_p) \in A^p : A \vDash \phi(v_1, \ldots, v_p)\}$. In other words, $\langle \phi, A \rangle$ is the probability that $\phi$ is satisfied in $A$ for a random (uniform independent) assignment of the free variables $x_1, \ldots, x_p$ to elements of $A$.

The above setting naturally extends to the case where a structure $A$ is equipped with a probability measure $\nu_A$ on its domain. In this case, we define the Stone pairing as

$$\langle \phi, A \rangle = \nu_A^{\otimes^p}(A),$$

where $\nu_A^{\otimes^p}$ stands for the product measure on $A^p$. In this paper we deal with finite structures endowed with a probability measure (which we briefly call structures for the sake of simplicity); when not defined, the probability measure considered on a finite structure is meant to be the uniform measure. The class of all the finite structures with signature $\sigma$ will be denoted by $\text{Rel}(\sigma)$.

In the following, we shall use the following convention (see Tables 1 and 2):
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Signification</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{Rel}(\sigma))</td>
<td>Set of all finite (\sigma)-structures (endowed with probability measure)</td>
</tr>
<tr>
<td>(A)</td>
<td>A structure</td>
</tr>
<tr>
<td>(\nu_A)</td>
<td>The domain of structure (A)</td>
</tr>
<tr>
<td>(X,Y)</td>
<td>Probability measure on the domain (A) of (A)</td>
</tr>
<tr>
<td>(X \cap Y, X \cup Y, X \setminus Y)</td>
<td>Subsets of (A)</td>
</tr>
<tr>
<td>(\phi(A))</td>
<td>Set of tuples satisfying (\phi) in (A)</td>
</tr>
<tr>
<td>(\langle \phi, A \rangle)</td>
<td>Stone pairing of (\phi) and (A)</td>
</tr>
<tr>
<td>(A[X])</td>
<td>Substructure of (A) induced by (X)</td>
</tr>
<tr>
<td>(A - X)</td>
<td>Substructure of (A) induced by (A \setminus X)</td>
</tr>
<tr>
<td>(N^d_A(X))</td>
<td>Closed (d)-neighborhood of subset (X) in (A)</td>
</tr>
<tr>
<td>(\partial_A(X))</td>
<td>Outer boundary of (X) in (A): (\partial_A(X) = N^d_A(X) \setminus X)</td>
</tr>
<tr>
<td>(A)</td>
<td>Sequence ((A_n)_{n \in \mathbb{N}}) of structures</td>
</tr>
<tr>
<td>(A)</td>
<td>Sequence ((A_n)_{n \in \mathbb{N}}) of the domains of structures in (A)</td>
</tr>
<tr>
<td>(X)</td>
<td>Sequence ((X_n)_{n \in \mathbb{N}}) of subsets, cluster</td>
</tr>
<tr>
<td>0</td>
<td>Sequence of empty sets: sequence (X), where (X_n = \emptyset)</td>
</tr>
<tr>
<td>(A_f)</td>
<td>Subsequence ((A_{f,n})_{n \in \mathbb{N}}) of (A)</td>
</tr>
<tr>
<td>(X_f)</td>
<td>Subsequence ((X_{f,n})_{n \in \mathbb{N}}) of (X)</td>
</tr>
<tr>
<td>(\nu_A(X))</td>
<td>Sequence ((\nu_A(X_n))_{n \in \mathbb{N}}) of measures of subsets</td>
</tr>
<tr>
<td>(X \cap Y, X \cup Y, X \setminus Y)</td>
<td>Sequences ((X_n \cap Y_n)<em>{n \in \mathbb{N}}, (X_n \cup Y_n)</em>{n \in \mathbb{N}}, (X_n \setminus Y_n)_{n \in \mathbb{N}})</td>
</tr>
<tr>
<td>(X \subseteq Y)</td>
<td>Pointwise sequence inclusion: (X \subseteq Y \iff (\forall n) X_n \subseteq Y_n)</td>
</tr>
<tr>
<td>(A[X])</td>
<td>Sequence ((A_n[X_n])_{n \in \mathbb{N}}) of induced substructures</td>
</tr>
<tr>
<td>(A - X)</td>
<td>Sequence ((A_n[A_n \setminus X_n])_{n \in \mathbb{N}}) of induced substructures</td>
</tr>
<tr>
<td>(N^d_A(X))</td>
<td>Sequence ((N^d_A(X_n))_{n \in \mathbb{N}}) of closed (d)-neighborhoods</td>
</tr>
<tr>
<td>(\partial_A X)</td>
<td>Sequence ((\partial_A(X_n))_{n \in \mathbb{N}}) of outer boundaries</td>
</tr>
<tr>
<td>(\phi(A))</td>
<td>Sequence ((\phi(A_n))_{n \in \mathbb{N}}) of satisfaction sets of (\phi)</td>
</tr>
<tr>
<td>(L_X(A))</td>
<td>Lifted sequence obtained by marking (X) in (A)</td>
</tr>
<tr>
<td>(\text{lim} A)</td>
<td>Limit of (A) (as an abstract object)</td>
</tr>
<tr>
<td>(\langle \phi, \text{lim} A \rangle)</td>
<td>Limit Stone pairing: (\langle \phi, \text{lim} A \rangle = \lim_{n \to \infty} \langle \phi, A_n \rangle)</td>
</tr>
</tbody>
</table>

**Table 1.** Main symbols and notations of this paper (part 1)

- Structures are denoted by boldface capital letters \(A\);
- Sets are denoted by plain roman capital letters \(X, Y\);
- Sequences of structures are denoted by boldface capital sans serif letter \(A = (A_n)_{n \in \mathbb{N}}\);
- Sequence of sets are denoted by plain capital sans serif letter \(X = (X_n)_{n \in \mathbb{N}}\).
Let $A$ be a $\sigma$-structure and let $X$ be a subset of the domain $A$ of $A$. If $v_A(X) > 0$ we define $A[X]$ as the substructure of $A$ induced by $X$ endowed with probability measure defined by $v_A[X](Y) = v_A(Y)/v_A(X)$ (for every $Y \subseteq X$). If $v_A(X) < 1$ we define $A - X = A[A \setminus X]$. We denote by Gaif$(A)$ the Gaifman graph of $A$. Recall that the Gaifman graph of $A$ is the graph with vertex set $A$ in which two vertices $x, y$ are adjacent if they belong together to some relation of $A$. The distance between two elements of a structure will always refer to the graph distance in the Gaifman graph of the structure.

For a subset $X \subseteq A$, the closed neighborhood of $X$ in $A$ is $N_A(X)$. Consequently the set of all elements of $A$ at distance at most $d$ from an element of $X$ is $N_A^d(X)$. The outer vertex boundary (or simply the outer boundary) of $X$ in $A$ is the set of vertices in $A \setminus X$ with at least one neighbor in $X$ [4]:

$$\partial_A X = N_A(X) \setminus X.$$ 

Note, in particular, that if $X$ is the domain of a union of connected components of $A$, then $\partial_A X = \emptyset$.

Furthermore, we extend all operations defined on structures and subsets to sequences coordinate-wise: The sequence $A$ has domain $A$ (meaning $A_n$ has domain $A_n$); for $X \subseteq A$ (meaning $X_n \subseteq A_n$) we denote by $A[X]$ the sequence $(A_n[X_n])_{n \in \mathbb{N}}$. 

<table>
<thead>
<tr>
<th>Introduced in Section 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_\sigma$</td>
</tr>
<tr>
<td>$P(S)$</td>
</tr>
<tr>
<td>$M_\sigma$</td>
</tr>
<tr>
<td>$\mu_A$</td>
</tr>
<tr>
<td>$\mu_{\lim A}$</td>
</tr>
<tr>
<td>$k(\phi)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Introduced in Section 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi \ominus \psi$</td>
</tr>
<tr>
<td>$\phi \ominus \psi$</td>
</tr>
<tr>
<td>$\phi \otimes \psi$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Introduced in Section 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \approx Y$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Introduced in Section 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X \nparallel Y$</td>
</tr>
</tbody>
</table>

Table 2. Main symbols and notations of this paper (part 2)
by $\partial_n X$ the sequence $(\partial_n X_n)_{n \in \mathbb{N}}$, by $X \subseteq Y$ the inclusions $X_n \subseteq Y_n$, by $\phi(A)$ the sequence of the sets $\phi(A_n)$, by $\nu_A(X)$ the sequence of the measures $\nu_{A_n}(X_n)$, etc. Also, for increasing $f : \mathbb{N} \rightarrow \mathbb{N}$ we denote by $A_f$ the subsequence $(A_{f(n)})_{n \in \mathbb{N}}$ of $A$ and by $X_f$ the subsequence $(X_{f(n)})_{n \in \mathbb{N}}$ of $X$. For instance, $A$ denotes a sequence of structures whose $n$th term is $A_n$, and $A$ denotes the sequence of the domains $A_n$ of the structures $A_n$.

**Definition 1.** For $\sigma$-structures $A_1, A_2, \ldots$ and not negative reals $\lambda_1, \lambda_2, \ldots$ with sum 1 we define $\sum_i \lambda_i A_i$ as the $\sigma$-structure $A$ obtained by endowing the disjoint union of the $\sigma$-structures $A_1, A_2, \ldots$ with the probability measure $\nu_A = \sum_i \lambda_i \nu_{A_i}$.

Note that although this allows us to define $\langle \phi, \sum_i \lambda_i A_i \rangle$, in general we have $\langle \phi, \sum_i \lambda_i A_i \rangle \neq \sum_i \lambda_i \langle \phi, A_i \rangle$. However, equality holds in the very particular case where $\phi$ is a local formula with a single free variable. When $\phi$ is a general local formula with $p$ free variables, it is possible to express $\langle \phi, \sum_i \lambda_i A_i \rangle$ as a polynomial of degree at most $p$ in terms of the form $\langle \phi_j, A_i \rangle$, for some strongly local formulas $\phi_j$ depending on $\phi$ (see Corollary 1).

To deal with “marking” we introduce the following notions of lift and shadow:

**Definition 2.** Let $\sigma \subseteq \sigma^+$ be countable signatures, let $A$ be a sequence of $\sigma$-structures, and let $B$ be a sequence of $\sigma^+$-structures.

The sequence $A$ is the shadow of the sequence $B$ if, for each $n \in \mathbb{N}$, the structure $A_n$ is the structure obtained from $B_n$ by “forgetting” about all relations not in $\sigma$. Conversely, the sequence $B$ is a lift of the sequence $A$ if $A$ is the shadow of $B$. The sequence $B$ is a conservative lift of the sequence $A$ if, for each $n \in \mathbb{N}$, the structures $A_n$ and $B_n$ have the same Gaifman graph.

Note that the terms lift and shadow (used here in a very general sense) are known in model theoretic literature as expansion and reduct [16]. Our choice of terminology is motivated here by the combinatorial context (see e.g. [15]).

In this paper, a lift of a sequence $A$ will usually be denoted by $L(A)$, with possibly adding some subscripts to differentiate different lifts of a same sequence. In particular, if $X$ is a sequence of subsets of $A$ (i.e. $X_n \subseteq A_n$) and $\sigma^+$ is the signature obtained from $\sigma$ by adding a single unary symbol $M$, we shall denote by $L_X(A)$ the lift of $A$ such that $M(L_X(A)) = X$.

For the benefit of the reader we included in Tables 1 and 2 the list of the main symbols and notations used throughout this paper.

3. Reduction from Local Formulas to Strongly Local Formulas

Recall that a first-order formula $\phi$ with free variables $x_1, \ldots, x_p$ is $r$-local if, for every structure $A$ and elements $v_1, \ldots, v_p \in A$ it holds

$$A \models \phi(v_1, \ldots, v_p) \iff A[N^r_A(\{v_1, \ldots, v_p\})] \models \phi(v_1, \ldots, v_p).$$

A formula is called local if it is $r$-local for some $r$. The set of all local first-order formulas (for given signature $\sigma$) is denoted by $\mathbf{FO}^\text{local}(\sigma)$. 


The following is the key definition.

**Definition 3.** A sequence $A = (A_n)_{n \in \mathbb{N}}$ of $\sigma$-structure is local-convergent if $(\langle \phi, A_n \rangle)_{n \in \mathbb{N}}$ convergences for every local formula $\phi$.

Note that this notion of convergence extends Benjamini-Schramm’s notion of local convergence:

**Fact 1 ([19]).** A sequence $(G_n)_{n \in \mathbb{N}}$ of graphs with maximum degree at most $D$ is local-convergent (in the sense of Benjamini-Schramm) if and only if it is local-convergent (in the sense of $\text{FO}^{\text{local}}$-convergence).

Indeed, by Gaifman locality theorem [10] (see also [16, Theorem 4.22]), for every local formula $\phi$ with $p$ free variables there exist $p$ formulas $\psi_1, \ldots, \psi_p$ with single free variable, such that for every graph $G$ with bounded degrees it holds

$$\langle \phi, G \rangle = \prod_{i=1}^{p} \langle \psi_i, G \rangle + o(1).$$

The main interest of our definition of local-convergence is that it does not need any restriction on the degrees. For general graphs (or regular hypergraphs), local convergence is stronger than the left convergence considered by [18, 8].

Before discussing the notion of local convergence in greater detail, we take time for few definitions.

**Fact 2.** Let $X, Y$ be subsets of the domain $A$ of a structure $A$, let $d$ be an integer and let $Z$ be any of $X \cap Y$, $X \cup Y$, $X \setminus Y$, $Y \setminus X$, $X \Delta Y$, and their complements in $A$. Then it holds

$$N^d_A(\partial_A Z) \subseteq N^{d+1}_A(\partial_A X) \cup N^{d+1}_A(\partial_A Y).$$

The notion of local formula can be strengthened by requiring that all the free variables are at bounded distance from each other. Precisely, a first-order formula $\phi$ with free variables $x_1, \ldots, x_p$ is strongly $r$-local if it is $r$-local and the following entailment holds:

$$\phi \vdash \bigwedge_{i=1}^{p} (\text{dist}(x_i, x_j) \leq r).$$

A formula is called strongly local if it is strongly $r$-local for some $r$.

We now introduce a notion of ”weak algebra” of formulas. In the following definition, a formula $\phi$ is packed if its free variables are $x_1, \ldots, x_p$ (for some $p \in \mathbb{N}$). For a formula $\phi$ with free variables $x_{i_1}, \ldots, x_{i_p}$ and an injection $i : \mathbb{N} \to \mathbb{N}$, $i(\phi)$ denotes the formula $\phi$ where all the occurrences of $x_j$ are replaced by $x_{i(j)}$. We denote by $\tau$ be the injection $i \mapsto i + 1$.

To formalize how strongly local formulas and local formulas are related we find useful to introduce the following notion of “weak algebra”.

Definition 4. Let $\mathcal{S}$ be a set of logical equivalence classes of first-order formulas. We say that $\mathcal{S}$ forms a weak algebra if it is closed under the following (partially defined) operations:

1. If $\phi, \psi \in \mathcal{S}$ and $\phi \land \psi = 0$ then $\phi \oplus \psi := \phi \lor \psi$ belongs to $\mathcal{S}$;
2. if $\phi, \psi \in \mathcal{S}$ and $\phi \rightarrow \psi$ then $\phi \ominus \psi := \phi \land \neg \psi$ belongs to $\mathcal{S}$;
3. if $\iota : \mathbb{N} \rightarrow \mathbb{N}$ is an injection and $\phi \in \mathcal{S}$ then $\phi^\iota \in \mathcal{S}$;
4. if $\phi, \psi \in \mathcal{S}$ are packed and $\phi$ has $p$ free variables, then $\phi \otimes \psi := \phi \land \tau^p(\psi)$ belongs to $\mathcal{S}$.

Note that for every $\phi, \psi$, we have:

- If $\phi \oplus \psi$ is defined then for every structure $A$ it holds
  $$(\phi \oplus \psi)(A) \equiv (\phi(A) \cup \psi(A))$$
  $$\langle \phi \oplus \psi, A \rangle = \langle \phi, A \rangle + \langle \psi, A \rangle$$

- If $\phi \ominus \psi$ is defined then for every structure $A$ it holds
  $$(\phi \ominus \psi)(A) \equiv (\phi(A) \setminus \psi(A))$$
  $$\langle \phi \ominus \psi, A \rangle = \langle \phi, A \rangle - \langle \psi, A \rangle$$

- If $\phi \otimes \psi$ is defined then for every structure $A$ it holds
  $$(\phi \otimes \psi)(A) = \phi(A) \times \psi(A)$$
  $$\langle \phi \otimes \psi, A \rangle = \langle \phi, A \rangle \cdot \langle \psi, A \rangle$$

Here, the equivalence $X \equiv Y$ (with respect to domain $A$) means, for $X \subseteq A^p$ and $Y \subseteq A^q$ that there exist a permutation $\iota$ of $[p + q]$ such that
$$\iota(X \times A^q) = Y \times A^p.$$ 

Note that all these operations preserve logical equivalence and thus the notion of weak algebra is well defined. Also, note that if $\phi \oplus \psi$ is defined then $\phi = (\phi \oplus \psi) \ominus \psi$.

Theorem 2. The smallest weak algebra containing all (logical equivalence classes of) strongly local formulas is the weak algebra of all (logical equivalence classes of) local formulas.

Proof. One direction is obvious (as local formulas form a weak algebra). For the other direction, consider an $r$-local formula $\phi$ with free variables $x_i$ for $i \in I$. Let $\mathcal{G}_I$ be the set of all graphs with vertex set $I$. Obviously it holds
$$\phi = \bigoplus_{F \in \mathcal{G}_I} \left( \bigwedge_{ij \in E(F)} (\text{dist}(x_i, x_j) \leq 2r) \land \bigwedge_{ij \not\in E(F)} (\text{dist}(x_i, x_j) > 2r) \land \phi \right).$$

It follows that we can restrict our attention to formulas of the form used in the above sum for some $F$. Moreover, we can assume that $I = [p]$ and that the vertex sets
$I_1, \ldots, I_k$ of the connected components $F_1, \ldots, F_k$ of $F$ are intervals of $[p]$. By locality property, we further assume that $\phi$ has the following form:

$$\phi = \bigwedge_{ij \in E(F)} (\text{dist}(x_i, x_j) \leq 2r) \land \bigwedge_{ij \notin E(F)} (\text{dist}(x_i, x_j) > 2r) \land \bigwedge_{z=1}^k \rho_z,$$

where $\rho_z$ is $r$-local with set of free variables $\{x_i : i \in I_z\}$.

We proceed by induction on $k$. If $k = 1$ then the formula is $2pr$-strongly local so the lemma holds. Assume that $k > 1$ and that the statement holds for less than $k$ connected components. For $1 \leq z \leq k$ define

$$\phi_z = \bigwedge_{ij \in E(F_z)} (\text{dist}(x_i, x_j) \leq 2r) \land \rho_z.$$

Then it holds

$$\bigotimes_{z=1}^k \phi_z = \bigoplus_{H \in \mathcal{G}'} \left( \bigwedge_{ij \in E(H)} (\text{dist}(x_i, x_j) \leq 2r) \land \bigwedge_{ij \notin E(H)} (\text{dist}(x_i, x_j) > 2r) \land \bigwedge_{z=1}^k \rho_z \right),$$

where $\mathcal{G}'$ is the set of all graphs $H$ with vertex set $[p]$ such that $H[I_z] = F_z$ for every $1 \leq z \leq k$. Note that $\mathcal{G}'$ contains exactly one graph with $k$ connected components, namely $F$, all the other ones containing strictly less than $k$ connected components. Thus, if we denote

$$\psi = \bigoplus_{H \in \mathcal{G}' \setminus \{F\}} \left( \bigwedge_{ij \in E(H)} (\text{dist}(x_i, x_j) \leq 2r) \land \bigwedge_{ij \notin E(H)} (\text{dist}(x_i, x_j) > 2r) \land \bigwedge_{z=1}^k \rho_z \right),$$

it holds

$$\phi = \bigotimes_{z=1}^k \phi_z \ominus \psi.$$

By induction, $\psi$ belongs to the weak algebra generated by strongly local formulas, hence so does $\phi$. \qed

We now take time for three important corollaries.

**Corollary 1.** For every $r$-local formula $\phi$ with $p$ free variables, there exist finitely many $(2pr)$-strongly local formulas $\phi_i$ ($1 \leq i \leq N$) and a polynomial $P \in \mathbb{Z}[X_1, \ldots, X_N]$ of degree at most $p$ such that for every structure $A$ it holds

$$\langle \phi, A \rangle = P(\langle \phi_1, A \rangle, \ldots, \langle \phi_N, A \rangle).$$

**Example 1.** Consider the formula $\phi(x_1, x_2)$, which asserts that $x_1$ and $x_2$ are at not equal or adjacent, and either $x_1$ and $x_2$ have the same degree or both have degree at least 4. This formula is clearly local, but not strongly local. It holds

$$\phi = \left( \bigoplus_{i=1}^4 \phi_i \otimes \phi_i \right) \ominus \psi,$$
where \( \phi_i(x_1) \) asserts that \( x_1 \) has degree \( i \) (for \( 1 \leq i \leq 3 \)) or degree at least 4 (if \( i = 4 \)), and \( \psi(x_1, x_2) \) is a strongly local formula, which asserts that \( x_1 \) and \( x_2 \) are equal or adjacent and either \( x_1 \) and \( x_2 \) have the same degree or both have degree at least 4. Hence for every graph \( G \) it holds

\[
\langle \phi, G \rangle = \sum_{i=1}^{4} \langle \phi_i, G \rangle^2 - \langle \psi, G \rangle.
\]

Note also the following corollary, which allows to check that first-order definable subsets of (the power of) a measurable structure are measurable (with respect to product measure) by reduction to sets definable by strongly local formulas.

**Corollary 2.** Assume \( A \) is an infinite structure, whose domain is a measurable space. If, for every strongly local formula \( \phi \) the set \( \phi(A) \) is measurable (with respect to product measure) then for every first-order formula \( \psi \) the set \( \psi(A) \) is measurable (with respect to product measure).

*Proof.* This is a direct consequence of Theorem 2 and Gaifman locality theorem [10]. □

**Corollary 3.** A sequence \( A \) is local-convergent if and only if it is strong-local-convergent.

### Part 2. Clustering Local Convergent Sequences

The notion of clustering we develop in this part is based on the stability of the convergence of a sequence when marking certain subsets of the domains of the structures in the sequence. This is expressed by the notion of lift introduced in Section 2.

Fig. 4 contains an example which we shall use as a running example to illustrate some of the concepts and results of this second part. This example illustrates a sample sequence \( G \), where \( G_n \) is built from an \( N \times N \) grid (where \( N \) is some increasing unbounded function of \( n \)) by adding locally some edges within regions \( A_1, \ldots, A_5 \), which have relative sizes and also distances between them as indicated. Within region \( A_1 \) (resp. \( A_2 \), edges are added randomly and independently with probability \( p_1 \) (resp. \( p_2 \)), within regions \( A_3 \) and \( A_4 \) edges are added randomly and independently with probability \( p_3 \), and within region \( A_5 \) are added the edges of a graph \( Z_n \), such that \( Z \) forms an expander sequence of regular graphs (for instance Margulis’ construction of an \( m \times m \) 8-regular expander). Considering a subsequence if necessary, we can assume that \( G \) is convergent, and that so are also \( G[A_i] \) (for \( i = 1, \ldots, 5 \)). Moreover, we assume that \( G[A_3] \) and \( G[A_4] \) have same limit (what is almost surely true).

### 4. Negligible Sets and Sequences

The following notion of negligible set corresponds intuitively to parts of the graph one can remove, without a great modification of the statistics of the graph.
Definition 5. Let \( A \) be a structure, let \( d \in \mathbb{N} \) and let \( \epsilon > 0 \). A subset \( X \subset A \) of elements of \( A \) is \((d, \epsilon)\)-negligible in \( A \) if

\[
v_A(N^d_A(X)) < \epsilon.
\]

The main property of \((d, \epsilon)\)-negligible sets is the following:
Lemma 1. Let $\phi \in \text{FO}^\text{local}_p$ be $r$-local with $r < d$, and let $X$ be a $(d, \epsilon)$-negligible set of a structure $A$. Then

$$|\langle \phi, A \rangle - \langle \phi, A - X \rangle| < 2\epsilon p.$$ \(\quad\)

Moreover, suppose $B$ is a structure with same domain as $A$, that $X$ is a $(d, \epsilon)$-negligible set of $B$, and that $A - X = B - X$. Then

$$|\langle \phi, A \rangle - \langle \phi, B \rangle| < \epsilon p.$$ \(\quad\)

Proof. We first prove the second inequality.

Consider the lift $L(A)$ (resp. $L(B)$) of $A$ (resp. $B$) where all elements in $N^d_X(A)$ (resp. $N^d_X(B)$) are marked with new unary relation $M$. Let $\psi(x_1, \ldots, x_p) := \bigvee M(x_i)$. Then it holds

$$0 \leq \langle \phi, L(A) \rangle - \langle \phi \land \neg \psi, L(A) \rangle < \langle \psi, L(A) \rangle < 1 - (1 - \epsilon)^p < \epsilon p$$

$$0 \leq \langle \phi, L(B) \rangle - \langle \phi \land \neg \psi, L(B) \rangle < \langle \psi, L(B) \rangle < \epsilon p$$

Thus, as $\phi(L(A)) = \phi(A), \phi(L(B)) = \phi(B)$ and $(\phi \land \neg \psi)(L(A)) = (\phi \land \neg \psi)(L(B))$ it holds

$$|\langle \phi, A \rangle - \langle \phi, B \rangle| < \epsilon p$$

For the second inequality, we have likewise

$$0 \leq \langle \phi, L(A) - X \rangle - \langle \phi \land \neg \psi, L(A) - X \rangle < \langle \psi, L(A) - X \rangle < \frac{\epsilon p}{\nu_A(A \setminus X)^p}$$

Moreover, as $\phi$ is $r$-local, it holds $(\phi \land \neg \psi)(L(A)) = (\phi \land \neg \psi)(L(A) - X)$ hence

$$\langle \phi \land \neg \psi, L(A) \rangle = \nu_A(A \setminus X)^p \langle \phi \land \neg \psi, L(A) - X \rangle.$$ 

Thus

$$|\langle \phi, L(A) \rangle - \nu_A(A \setminus X)^p \langle \phi, L(A) - X \rangle| < \epsilon p$$

Hence

$$|\langle \phi, L(A) \rangle - \langle \phi, L(A) - X \rangle|$$

$$\leq |\langle \phi, L(A) \rangle - \nu_A(A \setminus X)^p \langle \phi, L(A) - X \rangle| + 1 - \nu_A(A \setminus X)^p$$

$$< \epsilon p + 1 - (1 - \epsilon)^p$$

$$< 2\epsilon p.$$ \(\quad\)

hence the result, as $\langle \phi, L(A) \rangle = \langle \phi, A \rangle$ and $\langle \phi, L(A) - X \rangle = \langle \phi, A - X \rangle$. \(\Box\)

Definition 6. A $(d, \epsilon)$-fragmentation of a structure $A$ is a (at most) countable partition $(S, X_1, X_2, \ldots)$ of $A$ such that no element in $X_i$ has a neighbor in $X_j$ for $i \neq j$ and $S$ is $(d, \epsilon)$-negligible in $A$. \(\quad\)
Lemma 2. Assume \((S, X_1, X_2, \ldots)\) is a \((d, e)\)-fragmentation of \(A\) and let \(\phi\) be a strongly \(r\)-local formula \((r \leq d)\) with free variables \(x_1, \ldots, x_p\). Then
\[
\left| \langle \phi, A \rangle - \sum_{i \geq 1} \nu_{\langle \phi \rangle}(A[X_i]) \right| < 2pe.
\]

Proof. This follows from Lemma 1 and the fact that \(\phi(A - S)\) is the disjoint union of the structures \(\phi(A[X_i])\).

We now consider how the notion of \((d, e)\)-negligible subset of a structure allows to define negligible sequences of subsets and equivalence of sequences.

Definition 7. Let \(A\) be a local-convergent sequence of structures. A sequence \(X \subseteq A\) is negligible in \(A\) if
\[
\forall d \in \mathbb{N} : \limsup_{n \to \infty} \nu_{\langle A \rangle}(N^d_{\langle A \rangle}(X_n)) = 0,
\]
what we simply formulate as
\[
\forall d \in \mathbb{N} : \limsup_{n \to \infty} \nu_{\langle A \rangle}(N^d_{\langle A \rangle}(X)) = 0.
\]
Two sequences \(X\) and \(Y\) of subsets are equivalent in \(A\) (and we note \(X \approx Y\) if the sequence \(X \Delta Y = (X_n \Delta Y_n)_{n \in \mathbb{N}}\) is negligible in \(A\).

We denote by \(0\) the sequence of empty subsets. Hence \(X \approx 0\) is equivalent to the property that \(X\) is negligible.

We further define a partial order on sequences of subsets by \(X \leq Y\) if the sequence \(X \setminus Y = (X_n \setminus Y_n)_{n \in \mathbb{N}}\) is negligible in \(A\). Hence \(\leq\) has \(0\) for its minimum.

Two sequences \(A\) and \(B\) of structures are equivalent if there exists a negligible sequence \(X\) of \(A\) and a negligible sequence \(Y\) of \(B\) such that \(A_n - X_n\) is isomorphic to \(B_n - Y_n\) for every \(n \in \mathbb{N}\).

For an example of a negligible sequence of subsets and of equivalent sequences based on our running example, see Fig. 5.

The following lemma is a straightforward consequence of Lemma 1 but we think it nevertheless deserves to be explicitly stated.

Lemma 3. Let \(A\) and \(B\) be equivalent sequences of structures.

Then \(A\) is local-convergent if and only if \(B\) is local-convergent. In this case, they have the same limit.

5. CLUSTERS AND PRE-CLUSTERS

The notion of cluster of a local-convergent sequence we introduce now is a weak analog of the notion of union of connected components, or more precisely of the topological notion of “clopen subset”.

5. Clusters. In our setting, where clustering is performed on a local convergent sequence $A$, the term of “cluster”, which we will now define, will be used to qualify a sequence $X$ of sets, with $X_n \subseteq A_n$.

**Definition 8.** Let $A$ be a local-convergent sequence of structures.

A sequence $X \subseteq A$ is a cluster of $A$ if the following conditions hold:

1. the lifted sequence $L_X(A)$ obtained by marking set $X_n$ in $A_n$ by a new unary relation $M_X$ is local-convergent;
2. the sequence $\partial_A X$ is negligible in $A$.

Condition (1) can be seen as a continuity requirement for the subset selection. Condition (2) is stronger than the usual requirement that there are not too many connections leaving the cluster. We intuitively require that the (asymptotically arbitrarily large) ring around a cluster is very sparse zone. Note that no minimality nor connectivity assumption is made at this point.

We start our “cluster analysis” by means of the following notions (more will follow, see Fig. 3): A cluster $X$ is atomic if, for every cluster $Y$ of $A$ such that $Y \subseteq X$ either $Y \approx 0$ or $Y \approx X$; the cluster $X$ is strongly atomic if $X_f$ is an atomic cluster of $A_f$ for every increasing function $f : \mathbb{N} \to \mathbb{N}$. To the opposite, the cluster $X$ is a nebula if, for every increasing function $f : \mathbb{N} \to \mathbb{N}$, every atomic cluster $Y_f$ of $A_f$ with $Y_f \subseteq X_f$ is trivial. Finally, a cluster $X$ is universal for $A$ if $X$ is a cluster of every conservative lift of $A$.

**Figure 5.** On the left, our running example $G$ with a negligible set (in black); on the right, a sequence $H$ equivalent to $G$, obtained by removing the negligible set.
Lemma 4. Let \( \mathcal{X} \) be a cluster of \( \mathfrak{A} \) and let \( \mathcal{Y} \) be a sequence of subsets. Then \( \mathcal{X} \approx \mathcal{Y} \) in \( \mathfrak{A} \) if and only if \( \mathcal{Y} \) is a cluster of \( \mathfrak{A} \) and

\[
\limsup \nu_{\mathfrak{A}}(\mathcal{X} \Delta \mathcal{Y}) = 0.
\]

Proof. Assume \( \mathcal{Y} \) is a cluster and \( \limsup \nu_{\mathfrak{A}}(\mathcal{X} \Delta \mathcal{Y}) = 0 \). For every integer \( d \) it holds

\[
N_{A_n}^d(\mathcal{X}_n \Delta \mathcal{Y}_n) \subseteq (\mathcal{X}_n \Delta \mathcal{Y}_n) \cup N_{A_n}^d(\partial_{A_n}(\mathcal{X}_n \Delta \mathcal{Y}_n))
\]

\[
\subseteq (\mathcal{X}_n \Delta \mathcal{Y}_n) \cup N_{A_n}^d(\partial_{A_n}\mathcal{X}_n) \cup N_{A_n}^d(\partial_{A_n}\mathcal{Y}_n)
\]

Thus \( \mathcal{X} \Delta \mathcal{Y} \) is negligible in \( \mathfrak{A} \), that is \( \mathcal{X} \approx \mathcal{Y} \).

Conversely, assume \( \mathcal{X} \approx \mathcal{Y} \). Then obviously \( \limsup \nu_{\mathfrak{A}}(\mathcal{X} \Delta \mathcal{Y}) = 0 \). As \( \partial_{A_n}\mathcal{Y}_n \subseteq \partial_{A_n}\mathcal{X}_n \cup N_{A_n}(\mathcal{X}_n \Delta \mathcal{Y}_n) \), and as \( \mathcal{X} \Delta \mathcal{Y} \) is negligible since \( \mathcal{X} \approx \mathcal{Y} \), the sequence \( \partial_{A_n}\mathcal{Y} \) is negligible. Moreover, as \( L_{\mathcal{X}}(\mathfrak{A}) \approx L_{\mathcal{Y}}(\mathfrak{A}) \) (considering we use the same symbol for both lifts), we deduce that \( \mathcal{Y} \) is a cluster of \( \mathfrak{A} \).

\( \square \)

In particular, if \( \mathcal{X} \) is a cluster and \( \mathcal{X} \approx \mathcal{Y} \) then \( \mathcal{Y} \) is a cluster.

We have the following alternative characterization of clusters:

Lemma 5. Let \( \mathfrak{A} \) be a local-convergent sequence of structures.

A sequence \( \mathcal{X} \subseteq \mathfrak{A} \) is a cluster of \( \mathfrak{A} \) if either \( \mathcal{X} \approx 0 \) or the following conditions hold:

1. the sequence \( \mathfrak{A}[\mathcal{X}] \) is local-convergent;
2. the limit \( \lim \nu_{\mathfrak{A}}(\mathcal{X}) \) exists and is strictly positive;
3. the sequence \( \partial_{\mathfrak{A}}\mathcal{X} \) is negligible in \( \mathfrak{A} \).

Proof. Assume \( \mathcal{X} \) is a cluster of \( \mathfrak{A} \), and let \( \alpha = \langle M_{\mathcal{X}}, \lim L_{\mathcal{X}}(\mathfrak{A}) \rangle \). If \( \alpha = 0 \) then \( \mathcal{X} \) is negligible in \( \mathfrak{A} \) as for every \( d, n \in \mathbb{N} \) it holds \( N_{A_n}^d(\mathcal{X}_n) = \mathcal{X}_n \cup N_{A_n}^{d-1}(\partial_{A_n}\mathcal{X}_n) \) and thus

\[
\limsup \nu_{\mathfrak{A}}(N_{A}^d(\mathcal{X})) \leq \lim \nu_{\mathfrak{A}}(\mathcal{X}) + \limsup \nu_{\mathfrak{A}}(N_{A}^{d-1}(\partial_{\mathcal{X}}\mathcal{X})) = \alpha = 0.
\]

Otherwise, \( \alpha > 0 \). To every local formula \( \phi \) we associate the local formula \( \phi|M_{\mathcal{X}} \) conditioning every variables with \( M_{\mathcal{X}} \). Then it holds

\[
\langle \phi, A_n[X_n] \rangle = \langle \phi|M_{\mathcal{X}}, L_{\mathcal{X}}(A_n)[X_n] \rangle
\]

\[
= \alpha^{-p}\langle \phi|M_{\mathcal{X}}, L_{\mathcal{X}}(A_n) - \partial_{A_n}(X_n) \rangle + o(1)
\]

\[
= \alpha^{-p}\langle \phi|M_{\mathcal{X}}, L_{\mathcal{X}}(A_n) \rangle + o(1)
\]

It follows that the sequence \( \mathfrak{A}[\mathcal{X}] \) is local-convergent.

Conversely, let \( \mathcal{X} \) be a sequence satisfying the conditions of the lemma. Then either \( \mathcal{X} \approx 0 \) and \( \mathcal{X} \) is a cluster (according to Lemma 4), or \( \mathfrak{A}[\mathcal{X}] \) is local convergent, \( \lim \nu_{\mathfrak{A}}(\mathcal{X}) > 0 \), and \( \partial_{\mathfrak{A}}\mathcal{X} \approx 0 \). Then, denoting \( \alpha = \lim \nu_{\mathfrak{A}}(\mathcal{X}) \) it holds for every local formula \( \phi \) (with respect to the language of \( L_{\mathcal{X}}(\mathfrak{A}) \)), denoting \( \phi^+ \) (resp. \( \phi^- \)) the formula
where $M_{Y}$ is replaced by true (resp. false) it holds:

\[
\langle \phi, L_{X}(A_{n}) \rangle = \langle \phi, L_{X}(A_{n}) - \partial_{A_{n}} X_{n} \rangle + o(1)
\]

\[
= \alpha^{p} \langle \phi, L_{X}(A_{n})[X_{n}] \rangle + (1 - \alpha)^{p} \langle \phi, L_{X}(A_{n})[A_{n} \setminus X_{n} \setminus \partial_{A_{n}} X_{n}] \rangle + o(1)
\]

\[
= \alpha^{p} \langle \phi^{+}, L_{X}(A_{n})[X_{n}] \rangle + (1 - \alpha)^{p} \langle \phi^{-}, L_{X}(A_{n})[A_{n} \setminus X_{n} \setminus \partial_{A_{n}} X_{n}] \rangle + o(1)
\]

\[
= \alpha^{p} \langle \phi^{+}, A_{n}[X_{n}] \rangle + \langle \phi^{-}, A_{n} \rangle - \alpha^{p} \langle \phi^{-}, A_{n}[X_{n}] \rangle + o(1)
\]

Hence $L_{X}(A)$ is local convergent.

\[\square\]

**Definition 9.** Two clusters $X$ and $Y$ of a local-convergent sequence $A$ are *interweaving*, and we note $X \nparallel Y$ if every sequence $Z$ with $Z_{n} \in \{X_{n}, Y_{n}\}$ is a cluster of $A$.

For an example of interweaving clusters based on our running example, see Fig. 6. We characterize interweaving clusters in Lemma 6. An explicit example is provided by the sequence of graphs $(G_{n})_{n \in \mathbb{N}}$ where $G_{n}$ is the disjoint union of a path of length $n$ and a cycle of length $n$. The cluster formed by the paths and the cluster formed by the cycles are interweaving. (For another example, see Example 4.)

Interweaving clusters allow to build many new clusters by weaving (hence the name “interweaving”). Interweaving clusters have the following handy characterization:

**Lemma 6.** Let $X$ and $Y$ be two clusters of a local-convergent sequence $A$. The following are equivalent:

1. $X$ and $Y$ are interweaving;
2. $\lim L_{X}(A) = \lim L_{Y}(A)$;
3. either $X \approx Y \approx 0$ or the following two conditions hold:
   b. $\lim v_{A}(X) = \lim v_{A}(Y)$.

**Proof.** Let us prove (1)$\Rightarrow$(2)$\Rightarrow$(3)$\Rightarrow$(1).

(1)$\Rightarrow$(2): Let $Z_{n}$ be $X_{n}$ if $n$ is odd and $Y_{n}$ if $n$ is even. As $X \nparallel Y$, the sequence $Z$ is a cluster and (by considering the common subsequences) it holds $\lim L_{X}(A) = \lim L_{Z}(A) = \lim L_{Y}(A)$.

(2)$\Rightarrow$(3): Let $\alpha = \lim \langle M(x_{1}), L_{X_{n}}(A_{n}) \rangle$. Then either $\alpha = 0$ and $X \approx Y \approx 0$ or $\lim v_{A}(X) = \lim v_{A}(Y)\alpha > 0$. In the later case, for any local formula $\phi$, let $\tilde{\phi}$ be the formula where all the variables (free or bound) are constrained to belong to relation $M$. For sufficiently large $n$ (so that $\langle M(x_{1}), L_{X_{n}}(A_{n}) \rangle > 0$ and $\langle M(x_{1}), L_{Y_{n}}(A_{n}) \rangle > 0$) it holds

\[
\langle \phi, A_{n}[X_{n}] \rangle = \frac{\langle \tilde{\phi}, L_{X_{n}}(A_{n}) \rangle}{\langle M(x_{1}), L_{X_{n}}(A_{n}) \rangle} \quad \text{and} \quad \langle \phi, A_{n}[Y_{n}] \rangle = \frac{\langle \tilde{\phi}, L_{Y_{n}}(A_{n}) \rangle}{\langle M(x_{1}), L_{Y_{n}}(A_{n}) \rangle}.
\]


(3)⇒(1): If \( X \approx Y \approx 0 \) then obviously \( X \nmid Y \). Otherwise, consider arbitrary \( Z \) with \( Z_n \in \{ X_n, Y_n \} \). Assume for contradiction that \( A[Z] \) does not converge. Then we can extract two subsequences \( Z_f \) and \( Z_g \) such that \( \lim A_f[Z_f] \neq \lim A_g[Z_g] \). By taking further subsequences if necessary we can assume that \( Z_f \) and \( Z_g \) are each either a subsequence of \( X \) or \( Y \). As \( \lim A[X] = \lim A[Y] \) we get a contradiction, so \( A[Z] \) converges. Similarly, \( \nu(A[Z]) \) converges. As \( \partial A_n(Z_n) \subseteq \partial A_n(X_n) \cup \partial A_n(Y_n) \) and \( \partial A[X] \approx \partial A[Y] \approx 0 \) we get \( \partial A[Z] \approx 0 \). Altogether, this means that \( Z \) is a cluster of \( A \).

\[ \square \]

Obviously, interweaving is a limit for the possibility to track clusters in a local-convergent sequence. In some sense, interweaving clusters cannot be distinguished.

We now prove that the families of all clusters of a local convergent sequence is closed under complementation.

**Lemma 7.** Let \( A \) be a local-convergent sequence, and let \( X \) be a cluster of \( A \). Then \( Y = A \setminus X \) is a cluster of \( A \).

**Proof.** First notice that for every integer \( d \) it holds

\[ N^d_A(\partial A(Y)) \subseteq N^{d+1}_A(\partial A(X)) \]

thus \( \lim \nu(A(N^d_A(\partial A(Y)))) = 0 \), that is \( Y \approx 0 \). As \( L_Y(A) \) can be obtained from \( L_X(A) \) by taking for \( M_Y \) the negation of \( M_X \) it is clear that \( L_Y(A) \) is local convergent. \( \square \)

To the opposite, if \( X \) and \( Y \) are clusters, none of \( X \cap Y, \ Y \setminus X, \ X \Delta Y, \ X \cup Y \) and their complements needs to be a cluster. For that consider the following:

**Example 2.** Consider a local-convergent sequence \( E \) of connected expanders, where \( |E_n| = cn(1 + o(1)) \). Define the sequence \( A \) as follows:

\[
A_n = \begin{cases} 
E_{5n} \cup E_{6n} \cup E_{8n} & \text{if } n \text{ is odd} \\
E_{2n} \cup E_{3n} \cup E_{4n} \cup E_{10n} & \text{if } n \text{ is even}
\end{cases}
\]

Then it is easily checked that \( A \) is local convergent, and that the only clusters of \( A \) are (up to equivalence) \( 0, X, Y, A \setminus X, A \setminus Y, \) and \( A \), where

\[
X_n = \begin{cases} 
E_{5n} & \text{if } n \text{ is odd} \\
E_{2n} \cup E_{3n} & \text{if } n \text{ is even}
\end{cases}
\]

\[
Y_n = \begin{cases} 
E_{6n} & \text{if } n \text{ is odd} \\
E_{2n} \cup E_{4n} & \text{if } n \text{ is even}
\end{cases}
\]

Also notice that the graphs \( A_n \) could be made connected by linking connected components using paths of lengths \( \sqrt{n} \) without changing the conclusion.
Nevertheless, a simple necessary and sufficient condition for a family of clusters to generate a Boolean algebra of clusters can be given.

Lemma 8. Let $\mathbf{A}$ be a local convergent sequence, and let $C^1, \ldots, C', \ldots$ be countably many clusters of $\mathbf{A}$. Then the following conditions are equivalent:

1. The lifted sequence $L(\mathbf{A})$ defined by marking elements in $C^i$ by mark $M_i$ is local convergent;
2. The clusters $C^i$ generate a Boolean algebra of clusters, that is: every finite Boolean combination of $C^i$'s is a cluster;
3. Every finite intersection of $C^i$'s is a cluster.

Proof. We proceed by proving that (1) and (3) are both equivalent to (2).

That (1)$\Rightarrow$(2) is obvious as every finite Boolean combination of $C^i$'s is the set of solutions of the corresponding Boolean combination of $M_i$'s. Let us now prove (2)$\Rightarrow$(1). According to Corollary 3, in order to prove (1) it is sufficient to prove that for every strongly local formula $\phi$ (with some $p$ free variables) the sequence $\langle \phi, L(\mathbf{A}) \rangle$ converges. Let $N$ be the maximum index of a mark symbol used in $\phi$. For $I \subseteq [N]$ denote by $\phi^I$ the formula where every term $M_i(x)$ is replaced by $\text{true}$ if $i \in I$ and $\text{false}$ if $i \notin I$. Define $\theta_I$ as the formula $\bigwedge_{i \in I} M_i(x_1) \wedge \bigwedge_{i \notin I} \neg M_i(x_1)$. Let $S = \bigcup_{I \subseteq [N]} \partial_\mathbf{A} \theta_I(\mathbf{A})$. As each $\theta_I(\mathbf{A})$ defines a cluster, $S$ is negligible. Thus Then it holds

$$\langle \phi, L(\mathbf{A}_n) \rangle = \langle \phi, L(\mathbf{A}_n) - S \rangle + o(1)$$

$$= \sum_{I \subseteq [N]} \nu_{\mathbf{A}_n}(\theta_I(\mathbf{A}_n))^p \langle \phi^I, \mathbf{A}_n[\theta_I(\mathbf{A}_n)] \rangle + o(1)$$

thus $\langle \phi, L(\mathbf{A}_n) \rangle$ converges as $n \to \infty$. As this holds for every strongly local formula, we deduce that $L(\mathbf{A})$ is local convergent.

That (2)$\Rightarrow$(3) is trivial. Let us now prove (3)$\Rightarrow$(2). By following an easy induction and using the fact that the complement of a cluster is a cluster (Lemma 7) we reduce easily the implication to the following statement to be proved: if $X, Y$, and $X \cap Y$ are all clusters of $\mathbf{A}$ then so is $X \setminus Y$. To prove this, let $S = \partial_\mathbf{A} X \cup \partial_\mathbf{A} (X \cap Y)$. Then $S$ is negligible and thus for every strongly local formula $\phi$ (with $p$ free variables) it holds

$$\langle \phi, L_{X \setminus Y}(\mathbf{A}_n) \rangle = \langle \phi^0, \mathbf{A}_n \rangle + \nu_{\mathbf{A}_n}(X_n)^p \left( \langle \phi^1, \mathbf{A}_n[X_n] \rangle - \langle \phi^0, \mathbf{A}_n[X_n] \rangle \right)$$

$$+ \nu_{\mathbf{A}_n}(X_n \cap Y_n)^p \left( \langle \phi^0, \mathbf{A}_n[X_n \cap Y_n] \rangle - \langle \phi^1, \mathbf{A}_n[X_n \cap Y_n] \rangle \right) + o(1),$$

where $\phi^0$ (resp. $\phi^1$) stands for the formula obtained from $\phi$ by replacing each term of the form $M(x)$ by $\text{false}$ (resp. $\text{true}$). Hence $L_{X \setminus Y}(\mathbf{A}_n)$ is local convergent, and as $\partial_\mathbf{A}(X \setminus Y) \subseteq S$ is negligible, it follows that $X \setminus Y$ is a cluster. \hfill $\Box$

Definition 10. We say that two clusters $C_1$ and $C_2$ are

- weakly disjoint if $C_1 \Delta C_2 \approx 0$;
• disjoint if $C_1 \Delta C_2 = 0$;
• strongly disjoint if $(N_A(C_1) \cap C_2) \cup (C_1 \cap N_A(C_2)) = 0$.

With this definition, we have the following direct consequence of Lemma 8.

**Corollary 4.** Let $A$ be a local convergent sequence, and let $C^1, \ldots, C^i, \ldots$ be countably many pairwise weakly disjoint clusters of $A$.

Then the lifted sequence $L(A)$ defined by marking elements in $C^i$ by mark $M_i$ is local convergent.

The ultimate goal would be to extend Lemma 8 to the $\sigma$-algebra generated by the $C^i$’s. However, we do not expect this will be always the case, and we expect that some further conditions will be required.

For instance, in order to guarantee that countable unions will be clusters, it is natural to require that there is a negligible sequence including all the possible frontiers of countable Boolean combinations. Also, we shall need some “continuity” property for countable Boolean combinations. The simplest form for these conditions can be given when we further assume that the clusters $C^i$’s are pairwise weakly disjoint. In this case sufficient conditions for the property that the union of every countable subset of the $C^i$’s is a cluster are:

1. The sequence $\bigcup_{i} \partial_A C^i$ is negligible;
2. The clusters $C^i$ form a stable partition of $A$ in the sense that $\sum_{i} \lim \nu_A(C^i) = 1$.

**Lemma 9.** Let $A$ be a local convergent sequence, and let $C^1, \ldots, C^i, \ldots$ be countably many weakly disjoint clusters of $A$. Assume $\bigcup_{i} \partial_A C^i$ is negligible and $\sum_{i} \lim \nu_A(C^i) = 1$. Then for every $I \subseteq \mathbb{N}$, the sequence $\bigcup_{i \in I} C^i$ is a cluster. In other words, the collection of all unions of clusters among the $C^i$’s forms a $\sigma$-algebra of clusters.

**Proof.** Let $S = \bigcup_{i} \partial_A C^i$. Let $\phi$ be an $r$-local strongly local formula with $p$ free variables. Then for every positive real $\epsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$ it holds $\nu_A(N^{n+1}_A(S)) < \epsilon/8p$. Let $I \subseteq \mathbb{N}$ and let $L_I(A)$ be the sequence obtained from $A$ by marking elements of $\bigcup_{i \in I} C^i$ by a new mark $M$. Let $\psi$ be an $r$-local strongly local formula in the extended signature, and let $\psi^0$ (resp. $\psi^1$) be the formula obtained from $\psi$ by replacing each term of the form $M(x)$ by $false$ (resp. true). According to Lemma 2 we deduce that for every $n \geq n_0$ it holds

$$
\left| \langle \psi, L_I(A_n) \rangle - \sum_{i \in I} \nu_A(C^i_n) \langle \psi^0, A_n[C^i_n] \rangle - \sum_{i \in I} \nu_A(C^i_n) \langle \psi^1, A_n[C^i_n] \rangle \right| < \epsilon/4.
$$

As $\sum_{i \geq 1} \nu_A(C^i) = 1$ there exists $i_0 \in \mathbb{N}$ such that $\sum_{i \geq i_0} \nu_A(C^i) < \epsilon/8$ thus

$$
\sum_{i \in I, i > i_0} \left( \lim \nu_A(C^i) \right)^p \lim \langle \psi^0, A[C^i] \rangle + \sum_{i \in I, i > i_0} \left( \lim \nu_A(C^i) \right)^p \lim \langle \psi^1, A[C^i] \rangle < \epsilon/8.
$$

Moreover, the exists $n_1 \geq n_0$ such that for every $n \geq n_1$, every $1 \leq i \leq i_0$ and every $k \in \{0, 1\}$ it holds

$$
\left| \nu_A(C^i_n)^p \langle \psi^k, A_n[C^i_n] \rangle - \left( \lim \nu_A(C^i) \right)^p \lim \langle \psi^k, A[C^i] \rangle \right| < \epsilon/4i_0
$$

CLUSTER ANALYSIS OF LOCAL CONVERGENT SEQUENCES OF STRUCTURES 23
We deduce that

\[ \sum_{i > i_0} v_{A_n}(C_i) = 1 - \sum_{i = 1}^{i_0} v_{A_n}(C_i) < 1 - \sum_{i = 1}^{i_0} \lim v_A(C_i) + \epsilon/4 = \sum_{i > i_0} \lim v_A(C_i) + \epsilon/4 < 3\epsilon/8. \]

From this follows that

\[ \langle \psi, L_I(A_n) \rangle - \sum_{i \not\in I} (\lim v_A(C_i))^p \lim(\psi^0, A[C_i]) - \sum_{i \in I} (\lim v_A(C_i))^p \lim(\psi^1, A[C_i]) \rangle < \epsilon. \]

It follows that \( L_I(A) \) is local convergent. As \( \partial_A \left( \bigcup_{i \in I} C_i \right) \subseteq \bigcup_i \partial_A C_i \) is negligible by assumption, we deduce that \( \bigcup_{i \in I} C_i \) is a cluster. \( \Box \)

Note that when we consider the complete Boolean algebra generated by non weakly-disjoint clusters \( C_i \) the situation is less clear.

5.2. Universal Clusters. The next lemma states that the cluster of a sequence remain the same when marking a universal cluster.

Lemma 10. Let \( C \) be a universal cluster of a local convergent sequence \( A \), and let \( L_C(A) \) be the lift of \( A \) obtained by marking \( C \) by a new unary symbol \( M_C \).

Then, a sequence \( X \) is a cluster of \( A \) if and only if it is a cluster of \( L_C(A) \).

Proof. Of course, every cluster of \( L_C(A) \) is a cluster of \( A \).

Assume \( X \) is a cluster of \( A \). Then, by definition, the sequence \( L_X(A) \) is a local convergent lift of \( A \). As \( C \) is universal, it is a cluster of \( L_X(A) \) hence the sequence \( L_C(L_X(A)) \) is local convergent. As \( L_C(L_X(A)) = L_X(L_C(A)) \) we deduce that \( X \) is a cluster of \( L_C(A) \).

Also, marking a universal cluster preserves universal clusters (but new universal cluster may appear).

Remark 1. Let \( C \) be a universal cluster of a local convergent sequence \( A \), and let \( L_C(A) \) be the lift of \( A \) obtained by marking \( C \) by a new unary symbol \( M_C \).

Then, as every conservative lift of \( L_C(A) \) is a conservative lift of \( A \), it follows that every universal cluster of \( A \) is a universal cluster of \( L_C(A) \).

The universal clusters of \( A \) are of a particular interest, as they form (as we shall prove in the next two lemmas) a Boolean algebra of clusters preserved by conservative lifts, which includes all definable clusters of \( A \).

Lemma 11. Let \( A \) be a local convergent sequence and let \( \phi \) be a local formula with single free variable \( x_1 \).

Then the following conditions are equivalent:

1. \( \partial_A \phi(A) \approx 0 \);
(2) \( \phi(A) \) is a cluster of \( A \);
(3) \( \phi(A) \) is a universal cluster of \( A \).

**Proof.** If \( \phi(A) \) is a (universal) cluster of \( A \) then \( \partial_A \phi(A) \approx 0 \) (by definition of a cluster).

Conversely, if \( \partial_A \phi(A) \approx 0 \) then either \( \phi(A) \) is negligible (thus \( \phi(A) \) is a cluster) or \( A[\phi(A)] \) is local convergent: for every local formula \( \psi \) with free variables \( x_1, \ldots, x_p \), denoting \( \hat{\psi} \) the formula obtained by replacing terms \( \exists y \theta \) by \( (\exists y)(\phi(y) \land \theta) \) and terms \( (\forall y)\theta \) by \( (\forall y)(\phi(y) \rightarrow \theta) \) and denoting \( \widehat{\psi} \) the formula \( \widehat{\psi} \land \bigwedge_{i=1}^p \phi(x_i) \) we get

\[
\langle \psi, A \rangle_n[\phi(A_n)] = \langle \widehat{\psi}, A \rangle_n + o(1).
\]

It follows that \( \phi(A) \) is a cluster of \( A \). As condition (1) holds as well in every conservative lift of \( A \), it follows that \( \phi(A) \) is a universal cluster of \( A \) as well. \( \square \)

**Lemma 12.** Let \( A \) be a local convergent sequence. Then the equivalence classes of universal clusters of \( A \) form a Boolean algebra.

**Proof.** Let \( X, Y \) be universal clusters of \( A \), and let \( L(A) \) be a local convergent conservative lift of \( A \). Then the sequence \( L(L_X(L_Y(A))) \) is local convergent. If follows, by considering formulas \( \neg M_X, M_X \lor M_Y, \) and \( M_X \land M_Y \) that \( A \setminus X, X \cup Y \) and \( X \cap Y \) are clusters of \( L(L_X(L_Y(A))) \) hence of \( L(A) \). It follows that \( A \setminus X, X \cup Y \) and \( X \cap Y \) are universal clusters of \( A \). \( \square \)

### 5.3. Pre-Clusters.

**Definition 11.** A sequence \( X \) is a pre-cluster of \( A \) if \( X \approx 0 \) or if it holds

1. the sequence \( A[X] \) is local-convergent;
2. the limit \( \lim v_A(X) \) and is strictly positive;
3. for every integer \( d \) it holds

\[
\limsup v_A(N^d_A(X) \setminus X) = 0.
\]

The definition of pre-clusters of \( A \) is consistent with the notion of equivalence of sequence of subsets:

**Lemma 13.** Let \( X \) be a pre-cluster of \( A \) and let \( Y \approx X \) in \( A \). Then \( Y \) is a pre-cluster of \( A \).

**Proof.** That \( A[Y] \) is local-convergent follows from Lemma 3. Also, it is immediate that \( \lim v_A(Y) \) exists and that \( \lim v_A(Y) = \lim v_A(X) \).

Let \( Z = X \Delta Y \). By assumption, \( Z \) is negligible in \( A \).

Assume \( X \) is a pre-cluster. Let \( d \in \mathbb{N} \). Then

\[
N^d_A(Y) \setminus Y \subseteq (N^d_A(X) \cup N^d_A(Z)) \setminus (X \setminus Z) \\
\subseteq (N^d_A(X) \setminus X) \cup N^d_A(Z)
\]
It follows that \( \limsup \nu_A(N^d_A(Y) \setminus Y) = 0 \) hence \( Y \) is a pre-cluster.

**Lemma 14.** Every cluster is a pre-cluster.

**Proof.** This follows from the fact that \( N^d_A(X) \setminus X \subseteq N^d_A(\partial A_n X) \).

We now define a standard construction of a cluster from a pre-cluster.

**Definition 12.** Let \( X \) be a pre-cluster of a local-convergent sequence \( A \).

The *wrapping* of \( X \) in \( A \) is the sequence \( W \) obtained as follows:

For every \( n \in \mathbb{N} \), let \( D(n) \in \mathbb{N} \cup \{ \infty \} \) be the supremum of integers \( d \) such that for every \( n' \geq n \) it holds \( \nu_{A_n}(N^{2d+1}_{A_{n'}}(X_{n'}) \setminus X_{n'}) < 1/d \). Then we define \( W_n = N^{D(n)}_{A_n}(X_n) \).

Note that \( D(n) \) is non-decreasing and unbounded.

**Lemma 15.** For every pre-cluster \( X \) of \( A \), the wrapping \( W \) of \( X \) in \( A \) is (up to equivalence) the unique cluster such that \( X \subseteq W \) and

\[
\limsup \nu_A(W \setminus X) = 0.
\]

**Proof.** For every \( d \in \mathbb{N} \) there exists \( T(d) \) such that for every \( n \geq T(d) \) it holds \( \nu_{A_n}(N^{2d}_{A_n}(X_n) \setminus X_n) < 1/d \). For \( T(d) \leq n < T(d+1) \) we have \( W_n = N^d_{A_n}(X_n) \). Thus, for every \( d \in \mathbb{N} \) and every \( T(d') \leq n < T(d'+1) \) (with \( d' \geq d \)) it holds

\[
\nu_{A_n}(N^{d}_{A_n}(\partial A_n W_n)) \leq \nu_{A_n}(N^{2d'+1}_{A_n}(X_n) \setminus X_n) < 1/d'.
\]

Thus \( \partial A W \) is negligible in \( A \) hence \( W \) is a cluster of \( A \).

Moreover, for every \( n \geq N(d) \) it holds \( \nu_{A_n}(W_n \setminus X_n) < 1/d \).

Assume that a cluster \( Y \) of \( A \) as the same properties. Then

\[
\limsup \nu_A(W \setminus Y) \leq \limsup \nu_A(W \setminus X) + \limsup \nu_A(Y \setminus X) = 0.
\]

Hence, according to Lemma 4, \( W \) and \( Y \) are equivalent in \( A \).

5.4. *Expanding Clusters.* Here we introduce a sequential version of expansion property.

**Definition 13.** A structure \( A \) is \((d, \epsilon, \delta)-expanding\) if, for every \( X \subset A \) it holds

\[\epsilon < \nu_A(X) < 1 - \epsilon \implies \nu_A(N^d_A(X)) > (1 + \delta)\nu_A(X),\]

that is

\[
\inf \left\{ \frac{\nu_A(N^d_A(X) \setminus X)}{\nu_A(X)} : \epsilon < \nu_A(X) < 1 - \epsilon \right\} > \delta.
\]

Note that the left hand side of the above inequality is similar to the *magnification* introduced in [2], which is the isoperimetric constant \( h_{out} \) defined by

\[
h_{out} = \inf \left\{ \frac{|N_A(X) \setminus X|}{|X|} : 0 < \frac{|X|}{|A|} < 1/2 \right\}.
\]
Lemma 16. Let $0 < \epsilon < 1/6$ and let $A$ be a $(d, \epsilon, \delta)$-expanding structure. Then there exists a subset $Y \subseteq A$ of measure $\nu_A(Y) \leq \epsilon$ such that, denoting $A' = A - Y$, it holds
\[ \forall X \subseteq A' \quad \nu_{A'}(X) \leq 1/2 \implies \nu_{A'}(N_{A'}^d(X) \setminus X) \geq \delta \nu_{A'}(X). \]

Proof. Let $Y \subseteq A$ be maximal (for inclusion) with the property that $\nu_A(Y) \leq 1 - 2\epsilon$ and $\nu_A(N_A^d(Y) \setminus Y) < \delta \nu_A(Y)$. First note that $\nu_A(Y) \leq \epsilon$ as $A$ is $(d, \epsilon, \delta)$-expanding. Let $A' = A - Y$.

Assume for contradiction that there exists $Z \subset A'$ such that $\nu_{A'}(Z) \leq 1/2$ and $\nu_{A'}(N_{A'}^d(Z) \setminus Z) < \delta \nu_{A'}(Z)$. (Note that $\nu_A(Z) \leq \nu_{A'}(Z) \leq 1/2$.) As $\nu_{A'}$ is proportional to $\nu_A$ it also holds $\nu_A(N_A^d(Z) \setminus Z) < \delta \nu_A(Z)$. Moreover it obviously holds $N_A^d(Y \cup Z) \subseteq N_A^d(Y) \cup N_A^d(Z)$ thus
\[ \nu_A(N_A^d(Y \cup Z)) \leq \nu_A(N_A^d(Y)) + \nu_A(N_A^d(Z)) \]
\[ < (1 + \delta)(\nu_A(Y) + \nu_A(Z)) = (1 + \delta)(\nu_A(Y \cup Z)) \]
Hence $\nu_A(N_A^d(Y \cup Z) \setminus (Y \cup Z)) < \delta \nu_A(Y \cup Z)$. However $\nu_A(Y \cup Z) = \nu_A(Y) + \nu_A(Z) \leq \epsilon + 1/2 < 1 - 2\epsilon$, what contradicts the maximality of $Y$. \hfill \square

This lemma brings us even closer to the definition of the magnification. The main difference now stands in the existence of the parameter $d$. For graphs and $d = 2$, the sequence of stars shows that the concepts differ. Actually, for graphs, $(d, \epsilon, \delta)$-expansion means that the $d$th power of the graph (after deletion of a subset of vertices of measure at most $\epsilon$) has magnification at least $\delta$. In the very special (but standard) case of graphs with maximum degree at most $\Delta$ we recover the standard definition of expansion:

Lemma 17. Let $0 < \epsilon < 1/6$ and let $G$ be a $(d, \epsilon, \delta)$-expanding graph with degree at most $\Delta$. Then there $G$ has a subset $Y$ of size at most $\epsilon|G|$ such that $h_{out}(G - Y) \geq \delta/(\Delta - 1)^d$.

Proof. We consider the uniform probability measure on $G$. Then the lemma follows from Lemma 16 and the simple fact that if $G$ has maximum degree at most $\Delta$ then for every subset $X$ of vertices and for every integer $k \geq 1$ it holds $|N_{G}^{k+1}(X) \setminus N_{G}^{k}(X)| \leq (\Delta - 1)|N_{G}^{k}(X) \setminus N_{G}^{k-1}(X)|$, where we define $N_{G}^{0}(X) = X$. Hence $|N_{G}^{d}(X) \setminus X| \leq (1 + \cdots + (\Delta - 1)^{d-1})|N_{G}(X) \setminus X|$. \hfill \square

Definition 14. A local-convergent sequence $A$ is expanding if, for every $\epsilon > 0$ there exist $d, t \in \mathbb{N}$ and $\delta > 0$ such that every $A_n$ with $n \geq t$ is $(d, \epsilon, \delta)$-expanding.

A non-trivial cluster $X$ of $A$ is expanding of $A$ if $A[X]$ is expanding.

We have the following equivalent formulations of this concept:

Lemma 18. Let $X \not\approx \emptyset$ be a cluster of a local convergent sequence $A$. The following conditions are equivalent:

1. the sequence $A[X]$ is expanding;
(2) for every \( \epsilon > 0 \) there exists \( d, t \in \mathbb{N} \) such that for every \( Z \subseteq X \) with \( v_{A_n}(Z_n) > \epsilon v_{A_n}(X_n) \) and every \( n \geq t \) it holds
\[
v_{A_n}(N^d_{A_n}(Z_n)) > (1 - \epsilon)v_{A_n}(X_n);
\]
(3) the sequence \( X \) is a strongly atomic cluster of \( A \);
(4) for every \( \epsilon > 0 \) there exists no \( Y \subseteq X \) such that \( \partial_A Y \approx 0 \) and
\[
\epsilon < \liminf \frac{v_{A}(Y)}{v_{A}(X)} < \lim v_{A}(X) - \epsilon.
\]

**Proof.** First assume that \( A[X] \) is expanding, and assume for contradiction that \( X \) is not a strongly atomic cluster of \( A \). Then there exists some increasing function \( f : \mathbb{N} \to \mathbb{N} \) such that \( Y_f \) is a non-trivial cluster of \( A_f \), \( Y_f \subseteq X_f \) and \( Y_f \not\approx X_f \). Then \( \alpha = \lim v_{A_f}(Y_f)/v_{A_f}(X_f) \) is bounded away from 0 and 1. Thus there exists \( \delta > 0 \) and \( d \in \mathbb{N} \) such that
\[
\liminf \frac{v_{A_f}(N^d_{A_f}[X_f](Y_f))}{v_{A_f}(Y_f)} > 1 + \delta,
\]
what contradicts the property that \( Y_f \) is a cluster.

Now assume that \( X \) is a strongly atomic cluster of \( A \) and assume for contradiction that \( A[X] \) is not expanding. Then there exists \( \epsilon > 0 \) such that for every \( d \in \mathbb{N} \) it holds
\[
\liminf \inf_{n \to \infty} \frac{v_{A_n}(N^d_{A_n}[X_n](Y_n))}{v_{A_n}(Y_n)} = 1,
\]
where infimum is on subsets \( Y_n \subseteq X_n \) with \( \epsilon < v_{A_n}(Y_n)/v_{A_n}(X_n) < 1 - \epsilon \). We inductively construct an increasing function \( f : \mathbb{N} \to \mathbb{N} \) and subsets \( Y_{f(n)} \subseteq X_{f(n)} \) as follows: \( f(1) \) is the minimum integer \( n \) such that there exists \( Y_n \subseteq X_n \) with \( \epsilon < v_{A_n}(Y_n)/v_{A_n}(X_n) < 1 - \epsilon \) and \( v_{A_n}(N^d_{A_n}[X_n](Y_n)) < 2v_{A_n}(Y_n) \) and (for \( d \geq 1 \)) \( f(d+1) \) is the minimum integer \( n > f(d) \) such that there exists \( Y_n \subseteq X_n \) with \( \epsilon < v_{A_n}(Y_n)/v_{A_n}(X_n) < 1 - \epsilon \) and \( v_{A_n}(N^d_{A_n}[X_n](Y_n)) < d+2d^2_{A_n}[X_n](Y_n) \). It is easily checked that \( (Y_{f(n)}) \) is such that for every integer \( d \) it holds
\[
\limsup \frac{v_{A_f}(N^d_{A_f}(Y_f) \setminus Y_f)}{v_{A_f}(Y_f)} = 0.
\]
We can further consider a subsequence \( Y_{fog} \) of \( Y_f \) such that \( A_{fog}[Y_{fog}] \) is local convergent and \( v_{A_{fog}}(Y_{fog}) \) converges. It follows that \( Y_{fog} \) is a pre-cluster. Let \( \tilde{Y}_{fog} \) be the wrapping of \( Y_{fog} \) in \( A_{fog} \). Then \( \tilde{Y}_{fog} \) is a cluster, \( \tilde{Y}_{fog} \leq X_{fog} \) and \( \tilde{Y}_{fog} \not\approx X_{fog} \), what contradicts the assumption that \( X \) is a strongly atomic cluster.

A stronger form of expanding property is the non-dispersive property.

**Definition 15.** A local-convergent sequence \( A \) is *non-dispersive* if, for every \( \epsilon > 0 \) there exists \( d \in \mathbb{N} \) such that
\[
\liminf \sup_{n \to \infty} \frac{v_{A_n}(N^d_{A_n}(v_n))}{v_{A_n}(v_n)} > 1 - \epsilon.
\]
In other words, a sequence $A$ is non-dispersive if, for every $\varepsilon > 0$, $\varepsilon$-almost all elements of $A_n$ are included in some ball of radius at most $d$, for some fixed $d$.

**Definition 16.** A non-trivial cluster $X$ of $A$ is a globular cluster of $A$ if $A[X]$ is non-dispersive.

Every globular cluster is clearly strongly atomic, but the converse does not hold as witnessed, for instance, by sequence of expanders. The strongly atomic clusters that are not globular are called open clusters.

Opposite to globular clusters are residual clusters:

**Definition 17.** A cluster $X$ of $A$ is residual if for every $d \in \mathbb{N}$ it holds

$$\limsup_{n \to \infty} \sup_{v_n \in A_n} \text{sup} A_n^d (v_n) = 0.$$  

The case of bounded degree graphs is particularly interesting. Recall that a sequence $G$ of graphs is a vertex expander if there exists $\varepsilon > 0$ such that

$$\liminf h_{\text{out}}(G_n) \geq \varepsilon.$$  

For more information on expanders we refer the reader to [12].

**Lemma 19.** Let $G$ be a sequence of graphs with maximum degree at most $\Delta$ and let $C \not\equiv 0$ be a cluster of $G$. The following are equivalent:

- $C$ is a strongly atomic cluster;
- for every $\varepsilon > 0$ there exists $X \subseteq C$ such that for every $n \in \mathbb{N}$ it holds $|X_n| < \varepsilon |C_n|$ and $G[C \setminus X]$ is a vertex expander.

**Proof.** This is a direct consequence of Lemma 17. \qed

We now consider general expanding clusters (see Definition 14).

**Lemma 20.** Let $X$ be an expanding cluster of $A$, and let $Y$ be any cluster of $A$.

Then any convergent subsequence of $\left(\frac{\nu(X \cap Y)}{\nu(X)}\right)$ has limit either 0 or 1.

**Proof.** Let $Z = X \cap Y$. Assume there exists an increasing function $f : \mathbb{N} \to \mathbb{N}$ and a positive real $\alpha \in (0, 1)$ such that $\lim \nu_A f(X_f) / \nu_A f(X_f) = \alpha$.

According to Fact 2, for every integer $d \in \mathbb{N}$ it holds

$$N^d_{A_f(n)}(Z_f(n)) \subseteq N^{d+1}_{A_f(n)}(X_f(n)) \cup N^{d+1}_{A_f(n)}(Y_f(n)).$$  

It follows that $\partial_{A_f} Z_f$ is negligible in $A_f$.

By compactness, there exists an increasing function $g : \mathbb{N} \to \mathbb{N}$ such that $(A[Z])_{g \circ f}$ is local convergent. It follows that $Z_{g \circ f}$ is a cluster of $A_{g \circ f}$ which is neither equivalent to $0$ nor to $A_{g \circ f}$, thus $X$ is not strongly atomic, what contradicts the hypothesis, according to Lemma 18. \qed

**Lemma 21.** Let $X$ and $Y$ be expanding clusters of $A$.

Then
• either $X \cap Y \approx 0$ (i.e. $X$ and $Y$ are essentially disjoint);
• or $X \not\approx Y$ and then every $Z$ with $Z_n \in \{X_n, Y_n\}$ is an expanding cluster of $A$.

Proof. If $\nu_{A_n}(X_n \cap Y_n) = o(1)$ then $X$ and $Y$ are essentially disjoint (as $\partial_A(X \cap Y)$ is negligible in $A$ (see proof of Lemma 20).

Otherwise, there exists, according to Lemma 20, an increasing function $f : \mathbb{N} \to \mathbb{N}$ such that $X_f \approx Y_f$. It follows that $A_f[X_f]$ and $A_f[Y_f]$ (thus $A[X]$ and $A[Y]$) have the same local limit. Let $Z$ be such that $Z_n \in \{X_n, Y_n\}$. Then $A[Z]$ is local-convergent, $\partial_A Z$ is negligible in $A$, and $\lim \nu_A(Z)$ exists (and $\lim \nu_A(X) = \lim \nu_A(Y)$). It follows that $Z$ is a non-trivial cluster. Thus $X$ and $Y$ are interweaving (i.e. $X \not\approx Y$). That $Z$ is strongly atomic (hence expanding) follows from the hypothesis that both $X$ and $Y$ are expanding (hence strongly atomic): any cluster included in a subsequence of $Z$ has a subsequence which is a cluster included in a subsequence of $X$ or in a subsequence of $Y$. \qed

It is possible that a local-convergent sequence $A$ has arbitrarily many pairwise intersecting non-equivalent expanding clusters but not two essentially disjoint ones:

Example 3. Consider a local-convergent sequence $E$ of connected $d$-regular high-girth expanders with $|E_n| = cn(1 + o(1))$ (and uniform probability measure), for some constant $c > 1$. Let $A_n$ be defined as three copies of $E_n$ if $n$ is odd, and the union of $E_n$ and $E_{2n}$ if $n$ is even. Selecting a copy of $E_n$ into each $A_n$ leads to uncountably many pairwise intersecting non-equivalent expanding clusters. However, no two essentially disjoint expanding clusters exist in $A$. Note that we could have made $A_n$ connected by adding paths of length $\sqrt{n}$ linking the connected components without changing the conclusion.

More examples of clusters based on our running example are displayed on Fig. 6:

Example 4. The regions $A_1$ and $A_2$ being at bounded distance from each other, they belong to a same (strongly atomic, universal) globular cluster $C_1$. As $G[A_3]$ and $G[A_4]$ have the same limit, the globular clusters $C_4$ and $C_5$ are interweaving globular clusters. The cluster $C_2$ is non-atomic, the cluster $C_3$ is a (non-atomic residual) nebula, and the cluster $C_5$ is a (strongly atomic, residual) expanding cluster.

6. CLUSTERING AND THE CLUSTER COMB LEMMA

The notion of clustering intuitively covers the idea of partitioning the structures in a local convergent sequence as well as the limit into disjoint clusters.

Definition 18. Let $A$ be a local-convergent sequence of $\sigma$-structures. A lifted sequence $L(A)$ of $A$ obtained by extending the signature $\sigma$ into $\sigma^+$ by adding countably many unary relations $M_1, M_2, \ldots$ is a clustering of $A$ if, denoting

$$S = A \setminus \bigcup M_i(A)$$

the following conditions hold:
Figure 6. Some clusters of $G$.

(1) The sequence $L(A)$ is local convergent;
(2) The sequence $S$ is negligible and $\bigcup_i \partial A M_i(A) \subseteq S$;
(3) For every $n \in \mathbb{N}$, the non empty sets among $S_n, M_1(A_n), M_2(A_n), \ldots$ form a partition of $A_n$;
(4) The partition induced by the $M_i$’s is stable in the sense that
\[
\sum_i \lim \langle M_i, A \rangle = \lim \sum_i \langle M_i, A \rangle = 1.
\]

Remark 2. Conditions (1) and (2) imply that each sequence $M_i(A)$ is a cluster of $A$ hence a clustering of $A$ induce a “partition” into countably many disjoint clusters, and that the clusters defined by the marks $M_i$ are pairwise strongly disjoint (see Definition 10).

A simple idea to construct a clustering of a local convergent sequence $A$ is as follows: assume $A$ has a cluster $X_1 \approx 0$. Then let $\hat{A}_1$ be the lift of $A$ with $X_1$ marked $M_1$. If $X_1 \neq A$ look for a cluster $X_2 \approx 0$ of $\hat{A}_1$ disjoint from $M_1$ and mark it $M_2$, thus obtaining $\hat{A}_2$. Repeat the process until no cluster can be found. There are two main problems with this process:

- In general we do not obtain a clustering, as the obtained partition needs not to be stable and the global outer boundary $\bigcup_i \partial A M_i(A)$ needs not to be negligible;
- The partition is essentially not unique (and it is not clear which clusters of $A$ may appear simultaneously in the partition).
The first point is exemplified by the fact that we do not have the converse of Remark 2: partitioning into disjoint clusters do not define a clustering in general.

For instance, consider the following sequence of star forests.

**Example 5.** Consider the sequence $\mathbf{G}$ where $\mathbf{G}_n$ is the union of $2^n$ stars $\mathbf{H}_{n,1}, \ldots, \mathbf{H}_{n,2^n}$, where the $i$-th star $\mathbf{H}_{n,i}$ has order $2^{2^n}(2^{-i} + 2^{-n})/2$. Let $\mathbf{C}'$ be the sequence such that $C_n^i$ is the vertex set $H_{n,i}^n$ of the $i$th biggest connected component of $\mathbf{G}_n$ (or the empty subset if $i > 2^n$). It is easily checked that each $\mathbf{C}'$ is a cluster and that for each $n$ the (non-empty) subsets $C_n^i$ form a partition of $\mathbf{G}_n$. Assume that we mark each $\mathbf{C}'$ by mark $M_i$. Then, asymptotically, only one half of the vertices will be marked.

Nevertheless, we shall prove that the converse of Remark 2 is almost true. In order to do so, we consider countably many disjoint clusters $\mathbf{C}^1, \ldots, \mathbf{C}', \ldots$ of a local convergent sequence $\mathbf{A}$. For each $i \in \mathbb{N}$ we define

$$\lambda_i = \lim v_\mathbf{A}(C^i)$$

and

$$\lambda_0 = 1 - \sum_{i \geq 1} \lambda_i.$$ 

The next lemma shows how powerful the stability assumption (4) can be:

**Lemma 22.** Assume that there exists negligible $\mathbf{S} \supseteq \bigcup_i \partial_\mathbf{A} C^i$ and that it holds

$$\sum_i \lim v_\mathbf{A}(C^i) = \lim \sum_i v_\mathbf{A}(C^i).$$

Then $\mathbf{R} = \mathbf{A} \setminus \mathbf{S} \setminus \bigcup_i C^i$ is a cluster, and the lifted sequence $\mathbf{L}(\mathbf{A})$ obtained by marking $\mathbf{R}, \mathbf{C}^1, \mathbf{C}^2, \ldots$ by (say) marks $M_0, M_1, M_2, \ldots$ is a clustering of $\mathbf{A}$.

**Proof.** First note that (1) easily implies that $v_\mathbf{A}(C^i)$ converges to $(\lambda_i)_{i \in \mathbb{N}}$ in $\ell^p$-norm for $p \geq 1$. Let $\phi_1, \phi_2, \ldots$ be strongly $r$-local formulas with $p$ free variables in the language of $\mathbf{A}$. Then for any fixed $N \in \mathbb{N}$ it holds

$$\left| \left( \sum_{i \geq 1} v_\mathbf{A}(C^i_n)^p(\phi^i_n, \mathbf{A}_n[C^i_n]) \right) - \left( \sum_{i \geq 1} \lambda_i^p \lim(\phi^i_n, \mathbf{A}[C^i_n]) \right) \right| \leq \sum_{i \geq 1} |v_\mathbf{A}(C^i_n)^p - \lambda_i^p| + \sum_{i \geq 1} \lambda_i^p \left| \lim(\phi^i_n, \mathbf{A}[C^i_n]) - (\phi^i_n, \mathbf{A}_n[C^i_n]) \right|$$

$$\|v_\mathbf{A}(C^i) - (\lambda_i)_{i \in \mathbb{N}}\|_p + \sum_{i=1}^N \left| \lim(\phi^i_n, \mathbf{A}[C^i_n]) - (\phi^i_n, \mathbf{A}_n[C^i_n]) \right| + \sum_{i > N} \lambda_i^p.$$

It follows that

$$\lim_{n \to \infty} \sum_{i \geq 1} v_\mathbf{A}(C^i_n)^p(\phi^i_n, \mathbf{A}_n[C^i_n]) = \sum_{i \geq 1} \lambda_i^p \lim(\phi^i_n, \mathbf{A}[C^i_n]).$$
Let $\psi$ be a strongly $r$-local formula with $p$ free variables in the language of $\sigma^+ = \sigma \cup \{ M_0, M_1, M_2, \ldots \}$. For $\zeta$ non-negative integer, let $\psi^{\zeta}$ be the formula obtained from $\psi$ by replacing each term $M_i(t)$ by $\text{true}$ if $i = \zeta$ and $\text{false}$ otherwise. According to Lemma 2 it holds

$$
\langle \psi, L(A_n) \rangle = \nu_{A_n}(R_n)^p \langle \psi^0, A_n[R_n] \rangle + \sum_{i \geq 1} \nu_{A_n}(C_n^i)^p \langle \psi^i, A_n[C_n^i] \rangle,
$$

$$
\langle \psi^0, A_n \rangle = \nu_{A_n}(R_n)^p \langle \psi^0, A_n[R_n] \rangle + \sum_{i \geq 1} \nu_{A_n}(C_n^i)^p \langle \psi^0, A_n[C_n^i] \rangle.
$$

Thus, according to (2) it holds

$$
\lim_{n \to \infty} \sum_{i \geq 1} \nu_{A_n}(C_n^i)^p \langle \psi^0, A_n[C_n^i] \rangle = \sum_{i \geq 1} \lambda_i^p \lim \langle \psi^0, A[C_i] \rangle.
$$

Hence $\lim_{n \to \infty} \nu_{A_n}(R_n)^p \langle \psi^0, A_n[R_n] \rangle$ exists and

$$
\lim_{n \to \infty} \nu_{A_n}(R_n)^p \langle \psi^0, A_n[R_n] \rangle = \lim \langle \psi^0, A \rangle - \sum_{i \geq 1} \lambda_i^p \lim \langle \psi^0, A[C_i] \rangle.
$$

It follows that $\lim \langle \psi, L(A) \rangle$ exists and

$$
\lim \langle \psi, L(A) \rangle = \lim \langle \psi^0, A \rangle - \sum_{i \geq 1} \lambda_i^p \lim \langle \psi^0, A[C_i] \rangle + \sum_{i \geq 1} \lambda_i^p \lim \langle \psi^i, A[C_i] \rangle.
$$

Hence $L(A)$ is a clustering of $A$. \hfill \Box

To handle cases where (1) does not hold, we need to introduce the notion of clip:

**Definition 19.** A clip (for countably many disjoint clusters $C^1, \ldots, C', \ldots$ of a local-convergent sequence $A$) is a non-decreasing function $F : \mathbb{N} \to \mathbb{Z}^+$ such that $F \gg 1$ (i.e. $\lim_{n \to \infty} F(n) = \infty$) and such that for all integers $n \leq n'$ it holds

$$
\sum_{i=1}^{F(n)} \left| \nu_{A_{n'}}(C_{n'}^i) - \lambda_i \right| \leq \sum_{i>F(n)} \lambda_i.
$$

First, a few remarks are in order. The function $F : \mathbb{N} \to \mathbb{Z}^+$ defined by

$$
F(n) = \min \left( n, \max \left\{ t \leq n : \forall n' \geq n, \sum_{i=1}^t \left| \nu_{A_{n'}}(C_{n'}^i) - \lambda_i \right| \leq \sum_{i>t} \lambda_i \right\} \right)
$$

is a clip, as for $t = 0$ the inequality holds and as for every $k \in \mathbb{N}$ there exists $n \in \mathbb{N}$ such that $F(n) \geq \min(n, k)$ (as the left-hand side of the inequality tends to 0 as $n' \to \infty$). Thus clips always exist.

Secondly, remark that if $1 \ll G \leq F$ and $F$ is a clip then $G$ is a clip as well, as

$$
\sum_{i=1}^{G(n)} \left| \nu_{A_{n'}}(C_{n'}^i) - \lambda_i \right| \leq \sum_{i=1}^{F(n)} \left| \nu_{A_{n'}}(C_{n'}^i) - \lambda_i \right| \leq \sum_{i>F(n)} \lambda_i \leq \sum_{i>G(n)} \lambda_i.
$$
Lemma 23. Let $F$ be a clip. Then

$$\lim_{n \to \infty} \sum_{i=1}^{F(n)} \nu_{A_n}(C^i_n) = 1 - \lambda_0.$$  

Proof. It follows directly from the definition of a clip that for every integer $n$ it holds

$$\sum_{i=1}^{F(n)} \lambda_i - \sum_{i>F(n)} \lambda_i \leq \sum_{i=1}^{F(n)} \nu_{A_n}(C^i_n) \leq \sum_{i=1}^{F(n)} \lambda_i + \sum_{i>F(n)} \lambda_i.$$  

A $\lim_{n \to \infty} \sum_{i\geq F(n)} \lambda_i = 0$, we deduce that

$$\lim_{n \to \infty} \sum_{i=1}^{F(n)} \nu_{A_n}(C^i_n) = \sum_{i\geq 1} \lambda_i = 1 - \lambda_0.$$  

Given a clip $F$, we define $R$ by

$$R_n = A_n \setminus \bigcup_{i=1}^{F(n)} C^i_{n^i}.$$  

As for every integers $i$ and $d$ it holds

$$\lim \nu_A(N^d_A(\partial A C^i)) = 0$$

there exists a function $T : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ such that for every integers $i$, $d$ and $n \geq T(i, d)$ it holds

$$\nu_A(N^d_A(\partial A C^i_n)) \leq \frac{2^{-i}}{d}.$$  

Define

$$M(a) = \max_{1 \leq i \leq a} \max_{1 \leq d \leq a} T(i, j).$$  

Define also $G : \mathbb{N} \to \mathbb{Z}^+$ by

$$G(n) = \min(F(n), \max\{i \in \mathbb{N} : M(i) \leq n\}).$$  

Obviously, $1 \ll G \leq F$ thus $G$ is a clip. This clip has the following property:

Lemma 24. The sequence $S$ defined by

$$S_n = \bigcup_{i=1}^{G(n)} \partial A_n C^i_n$$

is negligible.
Proof. Let $d \in \mathbb{N}$. For sufficiently large $n$ it holds
\[
\nu_{A_n}(N^d_{A_n}(S_n)) \leq \nu_{A_n}(N^{G(n)}_{A_n}(S_n)) \\
\leq \sum_{i=1}^{G(n)} \nu_{A_n}(N^{G(n)}_{A_n}(\partial C^i_n)) \\
\leq \sum_{i=1}^{G(n)} \frac{2^{-i}}{G(n)} < \frac{1}{G(n)}.
\]
Hence
\[
\lim \nu_{A}(N^d_{A}(S)) = 0,
\]
that is: $S$ is negligible.

Define the subset sequences $D^i$ by
\[
D^i_n = \begin{cases} 
C^i_n & \text{if } n \geq G(i) \\
\emptyset & \text{otherwise}
\end{cases}
\]
and let $R$ be defined by $R_n = A_n \setminus \bigcup_{i \geq 1} D^i_n$.

Lemma 25. Either $\lambda_0 = 0$ and $R$ is negligible, or $\lambda_0 > 0$ and $R$ is a cluster.

Proof. Note that $S = \bigcup_{i \geq 1} \partial D^i \supseteq \partial R$.

In particular, $\partial R$ is negligible. According to Lemma 23, we have
\[
\lim_{n \to \infty} \sum_{i=1}^{G(n)} \nu_{A_n}(D^i_n) = 1 - \lambda_0,
\]
thus $\lim \nu_{A}(R) = \lambda_0$. Consider a strongly $r$-local formula $\phi$ with $p$ free variable. For every $\epsilon > 0$ there exists $n_0$ such that for every $n \geq n_0$ it holds $\nu_{A_n}(N^r_{A_n}(S_n)) < \epsilon$. It follows that
\[
\left| \langle \phi, A_n \rangle - \nu_{A_n}(R_n)^p \langle \phi, A_n[R_n] \rangle - \sum_{i \geq 1} \nu_{A_n}(D^i_n)^p \langle \phi, A_n[D^i_n] \rangle \right| < 3p\epsilon.
\]
Thus, if $\lambda_0 > 0$ we have
\[
\lim \langle \phi, A[R] \rangle = \frac{1}{\lambda_0^p} \lim_{n \to \infty} \left( \langle \phi, A_n \rangle - \sum_{i \geq 1} \lambda_i^p \langle \phi, A_n[D^i_n] \rangle \right) \\
= \frac{1}{\lambda_0^p} \left( \lim \langle \phi, A \rangle - \sum_{i \geq 1} \lambda_i^p \lim \langle \phi, A[D^i] \rangle \right).
\]
(Note that we can safely exchange limit and sum here because the partition into $R$ and the $D^i$’s is stable, see Lemma 22.) It follows that either $\lambda_0$ and $R$ is negligible, or $\lambda_0 > 0$ and $R$ is a cluster. \qed
Lemma 26 (Cluster Comb Lemma). Let $\mathbf{A}$ be a local convergent sequence of $\sigma$-structures, and let $C^1, \ldots, C^i, \ldots$ be countably many strongly disjoint clusters of $\mathbf{A}$.

Let $\sigma^+$ be the signature $\sigma$ augmented by unary relations $M_i (i \in \{0, 1, 2, \ldots \})$. Then there exist a clustering $\mathbf{A}^+$ of $\mathbf{A}$ with the property that for $i = 1, \ldots$, the marks $M_i$ comb the clusters $C^i$ in the sense that there exists a non decreasing function $G : \mathbb{N} \to \mathbb{N}$ with $g \gg 1$ with

$$M_i(\mathbf{A}_n) = \begin{cases} C^i_n & \text{if } n \geq G(i) \\ \emptyset & \text{otherwise} \end{cases}$$

(which implies $M_i(\mathbf{A}) \approx C^i$).

Proof. Denote $S = \mathbf{A} \setminus \bigcup_i M_i(\mathbf{A})$, $D^i = M_i(\mathbf{A})$ and $R = M_0(\mathbf{A})$.

Remark that we have the property that $\mathbf{A} \approx \mathbf{A} - S$, which is the disjoint union of $\mathbf{A}[R]$ and all the $\mathbf{A}[D^i]$. Mark vertices of $R$ by $M_0$, vertices of $D^i$ by $M_i$, and further mark vertices in $N_n$ by mark $M_S$. It is easily checked that the proportion of $A_n$ marked by some mark $M_i$ for $0 \leq i \leq k$ tends to $\sum_{i=0}^{k} \lambda_i$ as $n \to \infty$, and that this value tends to 1 as $k \to \infty$. Consider the signature $\sigma^+$ extended by these marks, and let $\mathbf{A}^+$ be the sequence of marked structures. Let $\phi$ be an $r$-local strongly local formula with $p$ free variables. Then

$$\langle \phi, \mathbf{A}^+_n \rangle = \lambda_0^p \langle \phi, \mathbf{A}^+_n[R_n] \rangle + \sum_{i \geq 1} \lambda_i^p \langle \phi, \mathbf{A}^+_{n[D^i_n]} \rangle.$$

Denote by $\phi_i$ the formula derived from $\phi$ by replacing each $M_i$ with $\mathbf{true}$ and every $M_j$ with $j \neq i$ with false. Notice that $\phi_i$ is an $r$-local strongly local formula in the language of the original signature $\sigma$, that $\phi(\mathbf{A}^+_n[R_n]) = \phi_0(\mathbf{A}^+_n[R_n])$ and that $\phi(\mathbf{A}^+_n[D^i_n]) = \phi_i(\mathbf{A}^+_n[D^i_n])$ (for $i \in \mathbb{N}$). Hence

$$\langle \phi, \mathbf{A}^+_n \rangle = \lambda_0^p \langle \phi_0, \mathbf{A}^+_n[R_n] \rangle + \sum_{i \geq 1} \lambda_i^p \langle \phi_i, \mathbf{A}^+_n[D^i_n] \rangle.$$

Thus $\mathbf{A}^+$ is a local convergent sequence.

Remark: If one only assumes that the clusters $C^i$ are pairwise weakly disjoint (meaning $C^i \Delta C^j$ is negligible if $i \neq j$) then we get the same conclusion, except that (4) is weakened to $M_i(\mathbf{A}) \approx C^i$. The idea is to define the clusters $Z^i = C^i \setminus \bigcup_{j<i} N_{\mathbf{A}}(C^j)$ that are strongly disjoints and equivalent to the original clusters.

7. The Clustering Problem

It is not clear which clusters of $\mathbf{A}$ can be “captured” in general from the only information available from local convergence, and whether it is possible to mark these clusters in a constructive way.
The answer to this question we get here is that we can always capture all the (countably many) globular clusters and that we can explicitly define the marking based on the knowledge of some of the limit Stone pairing and basic Fourier analysis, and a subtle cut method to handle the non-commutativity of countable sums and limits. This will be the motivation of the final part of this paper. As we shall see, this part demonstrates pleasing mathematical paradox: in order to achieve a more concrete result we first have to generalize.

Part 3. Effective Construction of the Globular Clusters

8. The Representation Theorem and Some Consequences

Let $B$ be the Lindenbaum–Tarski algebra defined by $\text{FO}^{\text{local}}(\sigma)$ and let $S_\sigma$ be the Stone dual of $B$, which is a Polish space, whose topology is generated by its clopen subsets. Recall that the duality of $B$ and $S_\sigma$ is expressed by the existence of a mapping $K$ from $\text{FO}^{\text{local}}(\sigma)$ to the family of all the clopen subsets of $S_\sigma$ such that $K(\phi \lor \psi) = K(\phi) \cup K(\psi)$, $K(\phi \land \psi) = K(\phi) \cap K(\psi)$, $K(\neg \phi) = S_\sigma \setminus K(\phi)$, and $K(\phi) = K(\psi)$ if and only if $\phi$ and $\psi$ are logically equivalent. For a local formula $\phi$, we further denote by $k(\phi)$ the indicator function of $K(\phi)$, which is obviously continuous on $S_\sigma$. Note that the $\sigma$-algebra of Borel subsets of $S_\sigma$ turns $S_\sigma$ into a standard Borel space.

The following representation theorem has been proved in [19] (in the case where finite structures are only considered with uniform measures). The extension to the general case (finite structures endowed with a probability measure) is easy, and we do not prove it here. (Note that the representation theorem proved in [19] is stated in a more general setting, where the considered formulas belong to a fixed fragment of first-order logic, which can possibly include all first-order formulas; the space $S_\sigma$ is then the Stone dual of the Lindenbaum-Tarski algebra generated by the considered fragment. One could ask whether results of this paper could be generalized to FO-clustering. This is not so, see Remark 11 below.)

**Theorem 3.** For every finite structure $M$ there is a unique probability measure $\mu_M$ on $S_\sigma$ such that for every local formula $\phi$ it holds

$$\langle \phi, M \rangle = \int_{S_\sigma} k(\phi) \, d\mu_M.$$ 

Moreover, for every two finite structures $M$ and $N$, it holds $\mu_M = \mu_N$ if and only if the structures obtained from $M$ and $N$ by removing connected components without non-zero weight elements are isomorphic as weighted structures.

Denote by $\mathfrak{M}_\sigma$ the closure of the space of all the probability measures $\mu_M$ for finite $M$ (with respect to weak topology).

Then, a sequence $A = (A_n)_{n \in \mathbb{N}}$ of finite $\sigma$-structure is local-convergent if and only if the sequence $(\mu_{A_n})_{n \in \mathbb{N}}$ of probability measures converges weakly, and then the limit probability measure is the unique probability measure $\mu_{\lim A}$ such that for every local
formula \( \phi \) it holds

\[
\int_{S_\sigma} k(\phi) \, d\mu_{\lim A} = \langle \phi, \lim A \rangle.
\]

Recall that a bounded sequence of positive finite measures \( \mu_n \) on \( S_\sigma \) converges weakly to the finite positive measure \( \mu \) if for any bounded continuous function \( f : S_\sigma \to \mathbb{R} \) it holds \( \int f \, d\mu_n \to \int f \, d\mu \). This is denoted by \( \mu_n \Rightarrow \mu \).

Thus for every continuous function \( f : S_\sigma \to \mathbb{R} \), and for every local convergent sequence \( A \) it holds \( \mu_{A_n} \Rightarrow \mu_{\lim A} \) and thus

\[
\int_{S_\sigma} f \, d\mu_{\lim A} = \lim_{n \to \infty} \int_{S_\sigma} f \, d\mu_{A_n}.
\]

(Note, however that (5) does not hold for general Borel functions \( f : S \to \mathbb{R} \).)

When considering random variables, one equally uses the terms convergence in distribution, weak convergence, or convergence in law. In our setting, we will use the term “weak convergence” when referring to convergence of probability measures on a Stone space, and we then use the notation \( \mu_n \Rightarrow \mu \); we will use the term “convergence in distribution” (or “convergence in law”) when referring to convergence random variables with values in \( \mathbb{R}^k \), and then we use the notation \( X_n \overset{d}{\to} X \). In this latter case, we use the term distribution (or law) of \( X \) for the related probability function on \( \mathbb{R}^k \). In the case of a (scalar) random variable \( X \), the distribution can be alternatively described by means of its cumulative distribution function \( F_X \) defined by \( F_X(x) = \mathbb{P}(X \leq x) \).

One of the important aspects of the study of local convergence is to determine (or even characterize) those parameters \( F \) that are local-continuous in the sense that if a sequence \( A = (A_n)_{n \in \mathbb{N}} \) of finite structures is local convergent then so is the sequence \( (F(A_n))_{n \in \mathbb{N}} \) of the associated parameters. Of course, every continuous real function \( f \in C(S_\sigma) \) defines a local continuous parameter \( A \mapsto \int_{S_\sigma} f \, d\mu_A \). But we shall make explicit some local continuous parameters that are not of this form. As we shall see such parameters will be of prime importance for clustering structures in a local convergent sequence.

**Definition 20.** Let \( A \) be a \( \sigma \)-structure and let \( \phi \) be a first order formula with free variables \( x_1, \ldots, x_p \) (with \( p \geq 1 \)). Denote by \( \phi^v(A) \) the set

\[
\phi^v(A) = \left\{ (u_1, \ldots, u_{p-1}) \in A^{p-1} : A \models \langle v, u_1, \ldots, u_{p-1} \rangle \right\}.
\]

The local Stone pairing of \( \phi \) and \( A \) at \( v \) is

\[
\langle \psi, A \rangle_v = \mathbb{P}(A \models \psi(v, X_2, \ldots, X_p)) = v_A^{\otimes(p-1)}(\phi^v(A))
\]

Hence if \( v_A(\{v\}) \neq 0 \) we get that the local Stone pairing of \( \phi \) and \( A \) at \( v \) is nothing but the conditional probability \( \mathbb{P}(A \models \psi(X_1, X_2, \ldots, X_p) \mid X_1 = v) \).
In our setting, every finite structure is considered as a probability space and thus the local Stone pairing of a formula $\phi$ and finite structure $A$ defines a random variable $\langle \phi, A \rangle \colon v \mapsto \langle \phi, A \rangle_v$.

The (admittedly technical) Lemma 27 will be the key tool for our estimation of clustering parameters. As it proceeds by means of Fourier analysis, we take time to recall some basics.

Given a random variable $X$ with values in $\mathbb{R}^k$ and law $P$, the characteristic function of $X$ or $P$ is

$$\gamma(t) = \mathbb{E}[e^{it \cdot X}] = \int e^{it \cdot x} dP(x) \quad \text{for every } t \in \mathbb{R}^k,$$

where $t \cdot x$ denotes the usual inner product of two vectors $x$ and $t$ in $\mathbb{R}^k$.

A standard Taylor expansion of $\mathbb{E}[e^{it \cdot X}]$ gives the following expression of the characteristic function as an infinite series:

$$\gamma(t) = \sum_{w_1 \geq 0} \cdots \sum_{w_k \geq 0} \mathbb{E}[X_1^{w_1} \cdots X_k^{w_k}] \frac{it_1^{w_1} \cdots t_k^{w_k}}{w_1! \cdots w_k!}.$$

A main property of characteristic functions is that they fully characterise distribution laws, and that they relate convergence in law of distributions to pointwise convergence of characteristic functions. Precisely, we have:

**Theorem 4** (Lévy’s continuity theorem). If $P_n$ are random laws on $\mathbb{R}^k$ whose characteristic functions $\gamma_n(t)$ converge for all $t$ to some $\gamma(t)$, where $f$ is continuous at 0 along each coordinate axis, then $P_n$ converges in law to a law $P$ with characteristic function $f$.

Note that there is a one-to-one correspondence between cumulative distribution functions and characteristic functions. If $X$ is a (scalar) random variable we have

**Theorem 5** (Lévy). If $\gamma$ is the characteristic function of a scalar random variable with cumulative distribution function $F$, then for two points $a < b$ such that $F$ is continuous at $a$ and $b$ it holds

$$F(b) - F(a) = \frac{1}{2\pi} \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \gamma(t) \, dt.$$

Moreover, if $a$ is an atom of $X$ (that is a discontinuity point of $F$) then

$$F(a) - F(a - 0) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} e^{-ita} \gamma(t) \, dt.$$

Note that this inversion theorem extends to the case of random vectors.

**Lemma 27** (Continuity of joint distribution of local Stone pairing). Let $\phi_1, \ldots, \phi_d$ be local formulas (with $p_1, \ldots, p_d$ free variables).
For \( \mu \in M_\sigma \) and \( t \in \mathbb{R}^d \) define
\[
\gamma(\mu, t) = \sum_{w_i \geq 0} \cdots \sum_{w_d \geq 0} \left( \int_{S_\sigma} k(\psi_w) \, d\mu \right) \prod_{j=1}^d \frac{(it_j)^{w_j}}{w_j!},
\]
where \( \psi_{(0, \ldots, 0)} \) is true statement, and for \( w \neq (0, \ldots, 0) \) we define
\[
\psi_w := \bigwedge_{i=1}^d \bigwedge_{j=1}^{w_i} \phi_i(x_{i,j,1}, \ldots, x_{i,j+p_i-1})
\]
with
\[
n_{i,j} = \left( \sum_{\ell=1}^{i-1} w_{\ell,p_\ell} \right) + (j-1)p_i + 1.
\]

Then, the following properties hold:

1. For every \( \mu \in M_\sigma \), the mapping \( t \mapsto \gamma(\mu, t) \) is the characteristic function of a \( d \)-dimensional random variable \( D(\mu) \);
2. The mapping \( \mu \mapsto D(\mu) \) is continuous in the sense that if \( \mu_n \) converges weakly to \( \mu \) then \( D(\mu_n) \) converges in distribution to \( D(\mu) \), that is:
   \[
   \mu_n \Rightarrow \mu \quad \Rightarrow \quad D(\mu_n) \xrightarrow{D} D(\mu);
   \]
3. For every finite structure \( A \) (with associated probability measure \( \mu_A \in M_\sigma \)) the \( d \)-dimensional random variable
   \[ D_A = (\langle \phi_1, A \rangle, \ldots, \langle \phi_d, A \rangle) \]
   has the same distribution as \( D(\mu_A) \).

Proof. We shall prove the three items in reverse order.

Let us prove (3). For any finite structure \( A \) and any vector \( w = (w_1, \ldots, w_d) \in \mathbb{N}^d \), let \( N = n_{d,w_d} + p_d - 1 \). Then it holds
\[
\psi_w(A) = \left\{ x \in A^N : (\forall i \in [d] \forall j \in [w_i]) (x_{i,j,1}, x_{i,j+1}, \ldots, x_{i,j+p_i-1}) \in \phi_i(A) \right\}
\]
\[
= \bigcup_{v \in A} \left\{ v \times x \in A^{N-1} : (\forall i \in [d] \forall j \in [w_i]) (x_{i,j}, \ldots, x_{i,j+p_i-2}) \in \phi_i^v(A) \right\}
\]
\[
= \bigcup_{v \in A} \underbrace{\{ v \} \times \phi_1^v(A) \times \cdots \times \phi_i^v(A) \times \cdots \times \phi_d(A) \times \cdots \times \phi_d(A)}_{w_1 \text{ times}}.
\]

Thus
\[
\langle \psi_w, A \rangle = \nu_{\psi_w}(A)
\]
\[
= \sum_{v \in A} v_{\psi_w}(\{ v \}) \left( v_{\phi_1}(A)^{w_1} \cdots (v_{\phi_d}(A)^{w_d}ight).
\]

Thus
\[
\langle \psi_w, A \rangle = \nu_{\psi_w}(A)
\]
\[
= \sum_{v \in A} v_{\psi_w}(\{ v \}) \left( v_{\phi_1}(A)^{w_1} \cdots (v_{\phi_d}(A)^{w_d}ight).
\]

Thus
\[
\langle \psi_w, A \rangle = \nu_{\psi_w}(A)
\]
\[
= \sum_{v \in A} v_{\psi_w}(\{ v \}) \left( v_{\phi_1}(A)^{w_1} \cdots (v_{\phi_d}(A)^{w_d}ight).
\]
If follows that the characteristic function $\gamma_A$ of $D_A$ is equal to:

$$\gamma_A(t) = \mathbb{E}[e^{itD_A}] = \sum_{w_1 \geq 0} \cdots \sum_{w_d \geq 0} \langle \psi_w, A \rangle \prod_{j=1}^d \frac{(it_j)^{w_j}}{w_j!} = \gamma(\mu_A, t).$$

(Note that as all the moments $\langle \psi_w, A \rangle$ are bounded by 1 the above series converges for every (complex) vector $t$.) As they have the same characteristic functions, the random variables $D_A$ and $D(\mu_A)$ have the same distribution.

Let us now prove (1) and (2). It is sufficient to consider the case where $\mu_n = \mu_{A_n}$ for some local convergent sequence $A$. As $\mu_n \Rightarrow \mu$, the functions $\gamma(\mu_n, t)$ converge pointwise to the function $\gamma(\mu, t)$. Moreover, $\gamma(\mu, t)$ is clearly continuous at $t = 0$ hence, according to Lévy’s continuity theorem, the random variables $D_{A_n}$ converge in distribution to a random variable $D$ with characteristic function $\gamma(\mu, t)$. \qed

Remark 3 (for an interested reader). The formula defining $\psi_w$ and the equality of $\langle \psi_w, A \rangle$ and $\mathbb{E}_v[\langle \phi_1, A \rangle_v^{w_1} \cdots \langle \phi_d, A \rangle_v^{w_d}]$ are generalization of the following simple fact: For a graph $G$ and a vertex $v$ of $G$, $\langle x_1 \sim x_2, G \rangle_v$ (where $\sim$ denotes adjacency) is the probability that a random vertex $x_1$ is adjacent to $x_1 = v$, that is $\deg(v) / |G|$, and $\langle x_1 \sim x_2, G \rangle$ is the average of $\langle x_1 \sim x_2, G \rangle_v$ over all vertices of $G$, that is $\langle x_1 \sim x_2, G \rangle = \mathbb{E}[\langle x_1 \sim x_2, G \rangle_v]$. Similarly, $\langle (x_1 \sim x_2) \land (x_1 \sim x_3), G \rangle_v$ is the probability that random $x_2$ is adjacent to $v$ and random $x_3$ is adjacent to $v$. As $x_2$ and $x_3$ are independent random vertices, this is nothing but the square of $\deg(v) / |G|$. Hence $\langle (x_1 \sim x_2) \land (x_1 \sim x_3), G \rangle = \langle (x_1 \sim x_2, G \rangle_v^2]$. The same way, for every integer $k$, it holds

$$\langle (x_1 \sim x_2) \land \cdots \land (x_1 \sim x_{k+1}), G \rangle = \langle (x_1 \sim x_2, G \rangle_v^k].$$

In this paper, we shall be interested in random variables that are a bit more complicated, but definable as a limit of local Stone pairing. In this context we will need the following complement to Lemma 27.

Lemma 28. Let $\mu \in \mathcal{M}_\alpha$ and let $(\phi_{\ell,1})_{\ell \in \mathbb{N}}, \ldots, (\phi_{\ell,d})_{\ell \in \mathbb{N}}$ be sequences of local formulas (with $p_1, \ldots, p_d$ free variables, respectively) such that for every integer $1 \leq i \leq d$ it holds

$$\phi_{1,i} \rightarrow \phi_{2,i} \rightarrow \cdots \rightarrow \phi_{\ell,i} \rightarrow \cdots$$

(where $\rightarrow$ stands for logical implication).

Let $D_\ell(\mu)$ be a $d$-dimensional random variable with characteristic function $\gamma(\mu, t)$, which is the function associated to $\phi_{\ell,1}, \ldots, \phi_{\ell,d}$ as in Lemma 27.

Then, as $\ell \rightarrow \infty$, the random variables $D_\ell(\mu)$ converge in distribution to a random variable $D_\infty(\mu)$, whose characteristic function $\gamma_\infty(\mu, t)$ is the pointwise limit of the functions $\gamma(\mu, t)$.

Proof. Let

$$\psi_{\ell,w} := \bigwedge_{i=1}^d \bigwedge_{j=1}^{w_i} \phi_{\ell,i}(x_1, x_{n_{i,j}+1}, \ldots, x_{n_{i,j}+p_i-1})$$
(as in Lemma 27).
For each vector $\mathbf{w} \in \mathbb{N}^d$ the sequence $\left( \int_S k(\psi_{\ell, \mathbf{w}}) \, d\mu \right)_{\ell \in \mathbb{N}}$ is non-decreasing and bounded by 1 hence converging. It follows that the functions $\gamma_{\ell}(\mu, \mathbf{t})$ converge point-wise as $\ell \to \infty$ to
\[
\gamma_{\infty}(\mu, \mathbf{t}) = \sum_{w_1 \geq 0} \cdots \sum_{w_d \geq 0} \left( \lim_{\ell \to \infty} \int_S k(\psi_{\ell, \mathbf{w}}) \, d\mu \right) \prod_{j=1}^d \frac{(\ell t_j)^{w_j}}{w_j!},
\]
which is continuous at $\mathbf{t} = \mathbf{0}$. Thus the theorem follows from Lévy’s continuity theorem.

Note that if $\mathbf{A}$ is a local convergent sequence of finite structures, it holds
\[
\mathbf{D}_{\ell}(\mu_{\mathbf{A}_n}) \overset{D}{\to} \mathbf{D}_{\ell}(\mu_{\lim \mathbf{A}}) \quad \text{as } n \to \infty
\]
\[
\mathbf{D}_{\ell}(\mu_{\lim \mathbf{A}}) \overset{D}{\to} \mathbf{D}_{\infty}(\mu_{\lim \mathbf{A}}) \quad \text{as } \ell \to \infty
\]
However, although $\mathbf{D}_{\ell}(\mu_{\mathbf{A}_n}) \overset{D}{\to} \mathbf{D}_{\infty}(\mu_{\mathbf{A}_n})$ as $\ell \to \infty$, it is not true in general that $\mathbf{D}_{\infty}(\mu_{\mathbf{A}_n})$ converges in distribution to $\mathbf{D}_{\infty}(\mu_{\lim \mathbf{A}})$ as $n \to \infty$.

**Definition 21.** Assume $\mathbf{A}$ be a $\sigma$-structures. The 1-point random lift distribution of $\mathbf{A}$ is the probability distribution over (isomorphism classes of) $\sigma^*$-structures (where $\sigma^*$ is the signature obtained from $\sigma$ by adding a unary symbol $M_1$), corresponding to the marking a random elements $X_1$ of $\mathbf{A}$, drawn from $\mathbf{A}$ according to probability distribution $\nu_{\mathbf{A}}$. We denote by $\Pi$ the map from the space $\text{Rel}(\sigma)$ of isomorphism classes of finite $\sigma$-structures (with domain endowed with a probability measure) to the space $P(\text{Rel}(\sigma))$ of probability distributions over $\text{Rel}(\sigma)$, which maps a $\sigma$-structure $\mathbf{A}$ to its 1-point random lift distribution $\Pi(\mathbf{A})$.

Recall that in the context of structures with a domain endowed with a probability measure, the notion of isomorphism is more involved than standard isomorphism of structures with no associated probability measure.

Let $\mathbf{A}$ and $\mathbf{B}$ be $\sigma$-structures, and let $N_A$ (resp. $N_B$) be the union of all the connected components of $\mathbf{A}$ (resp. $\mathbf{B}$) without any element of positive measure. Then $\mathbf{A}$ and $\mathbf{B}$ are isomorphic if there exists a bijective mapping $f : A \setminus N_A \to B \setminus N_B$ preserving the measure (i.e. such that $\nu_{\mathbf{A}} = \nu_{\mathbf{B}} \circ f$) and all the relations both ways (i.e. $\mathbf{A} \models R(u_1, \ldots, u_n) \iff \mathbf{B} \models R(f(u_1), \ldots, f(u_n))$).

**Remark 4.** The 1-point random lift corresponds to marking a vertex by $M_1$. Thus the obtained structure is a “rooted” structure. We choose this terminology in view of generalization to multiple and iterated random rooting.

The space $\text{Rel}(\sigma)$, endowed with topology defined by local convergence, can be identified (via the continuous injection $i^\sigma : \mathbf{A} \mapsto \mu_{\mathbf{A}}$ of the representation theorem) to an open subspace of the Polish space $P(S_{\sigma})$, the space of all probability measures on $S_{\sigma}$ (with weak-* topology). We denote by $\mathcal{M}_\sigma$ the closure of $i^\sigma(\text{Rel}(\sigma))$. Similarly,
the space $\text{Rel}(\sigma^*)$ can be identified via injection $i^*$ to an open subspace of $P(S_{\sigma^*})$ with closure $\mathcal{M}_{\sigma^*}$. The pushforward $i^*_*: P(\text{Rel}(\sigma)) \to P(\mathcal{M}_{\sigma^*})$ of $i^*$, defined by

$$i^*_*(\zeta) = \zeta \circ (i^*)^{-1}$$

is a continuous injection from $P(\text{Rel}(\sigma))$ to $P(\mathcal{M}_{\sigma^*})$.

The following result makes possible to transfer results about unrooted structures to 1-point random lifts (i.e. randomly rooted structures). It is a non-trivial refining of Representation Theorem 3 and it is the main result of this section.

**Theorem 6** (1-point random lift theorem). There exists a (unique) continuous function $\tilde{\Pi}: \mathcal{M}_{\sigma} \to P(\mathcal{M}_{\sigma^*})$ such that the following diagram commutes:

$$\begin{array}{ccc}
\text{Rel}(\sigma) & \xrightarrow{\Pi} & P(\text{Rel}(\sigma^*)) \\
\downarrow i^* & & \downarrow i^*_* \\
\mathcal{M}_{\sigma} & \xrightarrow{\tilde{\Pi}} & P(\mathcal{M}_{\sigma^*})
\end{array}$$

**Proof.** Consider an enumeration $\phi_1, \ldots, \phi_d, \ldots$ of local formulas with respect to signature $\sigma^*$. To each formula $\phi_i$ with $p \geq 0$ free variables we associate the local formula $\psi_i$ (with respect to signature $\sigma$) with $p+1$ free variables by replacing each free variable $x_i$ by $x_{i+1}$, and then each term $M_1(t)$ by the term $t = x_1$. Consider $\sigma^*$-structures $A^+$ obtained by marking a single element $v \in A$ in a $\sigma$-structure $A$. Then it holds

$$\langle \phi_i, A^+ \rangle = \langle \psi_i, A \rangle_v.$$

In order to prove Theorem 6, it is sufficient to prove that if $(A_n)_{n \in \mathbb{N}}$ is a local convergent sequence, then the measures $\rho_{\sigma^*} \circ \Pi(A_n)$ converge weakly. This is sufficient as for every $\mu \in \mathcal{M}_{\sigma}$ we can then define $\tilde{\Pi}(\mu)$ as the weak limit of the measures $\rho_{\sigma^*} \circ \Pi(A_n)$, where $(A_n)_{n \in \mathbb{N}}$ is any sequence of finite $\sigma$-structures such that $\mu_{A_n} \Rightarrow \mu$. This proves Theorem 6.

Thus let $(A_n)_{n \in \mathbb{N}}$ be a local convergent sequence and let $\zeta_n = \rho_{\sigma^*} \circ \Pi(A_n)$. The topology of $\mathcal{M}_{\sigma^*}$ can be metrized by means of the following metric: for $\mu_1, \mu_2 \in \mathcal{M}_{\sigma^*}$, we define the distance $d(\mu_1, \mu_2)$ by

$$d(\mu_1, \mu_2) = \inf \left\{ \varepsilon > 0 : \forall 1 \leq i \leq 1/\varepsilon \left| \int_{S_{\sigma^*}} k(\phi_i) \, d\mu_1 - \int_{S_{\sigma^*}} k(\phi_i) \, d\mu_2 \right| \leq \varepsilon \right\}.$$

Let $F: \mathcal{M}_{\sigma^*} \to [0, 1]$ be continuous, and let $\varepsilon > 0$. As $\mathcal{M}_{\sigma^*}$ is compact there is $\alpha > 0$ such that for every $\mu_1, \mu_2 \in \mathcal{M}_{\sigma^*}$ it holds

$$d(\mu_1, \mu_2) \leq \alpha \quad \Rightarrow \quad |F(\mu_1) - F(\mu_2)| \leq \varepsilon.$$
Let $d = \lceil 1/\alpha \rceil$. Consider the continuous map $p : \mathcal{M}_\sigma \to [0, 1]^d$ defined by

$$p(\mu) = \left( \int_{S_\sigma} k(\phi_1) \, d\mu, \ldots, \int_{S_\sigma} k(\phi_d) \, d\mu \right).$$

Consider a partition $B_1, \ldots, B_{d^d}$ of $[0, 1]^d$ into $d^d$ boxes of side $1/d$ ("Rubik cube type partition"). Let $C_j = p^{-1}(B_j)$, and let $t_j = F(\mu_j)$ for an arbitrary (fixed) choice of $\mu_j \in C_j$. According to Lemma 27, the sequence of tuples $((\psi_1, A_n), \ldots, (\psi_d, A_n))$ converges in distribution. Thus for every box $C_j$ the value

$$\int_{C_j} d\zeta_n(\mu) = \Pr[\mu \in C_j] \quad (\mu \text{ dist. wrt } \zeta_n)$$

$$= \Pr[((\phi_1, A_n^+), \ldots, (\phi_d, A_n^+)) \in B_j] \quad (A_n^+ \text{ dist. wrt } \Pi(A_n))$$

$$= \Pr[((\psi_1, A_n), \ldots, (\psi_d, A_n)) \in B_j] \quad (v \text{ dist. wrt } v_{A_n})$$

converges as $n \to \infty$. For every $1 \leq j \leq d^d$ and for every $\mu_1, \mu_2 \in C_j$ it holds $d(\mu_1, \mu_2) \leq 1/d$ hence for every $\mu \in C_j$ it holds $|F(\mu_j) - t_j| \leq \epsilon$. Thus it holds

$$\left| \int_{C_j} F(\mu) \, d\zeta_n(\mu) - t_j \int_{C_j} d\zeta_n(\mu) \right| \leq \epsilon \int_{C_j} d\zeta_n(\mu).$$

As the sets $C_j$ form a partition of $S_\sigma$, and as $\zeta_n$ is a probability measure, we get

$$\left| \int_{S_\sigma} F(\mu) \, d\zeta_n(\mu) - \sum_{j=1}^{d^d} t_j \int_{C_j} d\zeta_n(\mu) \right| \leq \epsilon.$$

Hence for sufficiently large $n$, $\int_{S_\sigma} F(\mu) \, d\zeta_n(\mu)$ concentrates in an interval of size at most $2\epsilon$. By letting $\epsilon \to 0$ we conclude that $\int_{S_\sigma} F(\mu) \, d\zeta_n(\mu)$ converges, hence (as this holds for every continuous function $F$) that $\zeta_n$ is weakly convergent.

\textbf{Remark 5.} Actually, along the same lines we could prove more: for the linear operator $\hat{\Pi} : P(\text{Rel}(\sigma)) \to P(\text{Rel}(\sigma^*))$ defined by

$$\hat{\Pi}(\zeta) = \int_{\text{Rel}(\sigma)} \Pi(A) \, d\zeta(A),$$

there exists a (unique) continuous linear map $\tilde{\Pi}$ such that the following diagram commutes:

$$\begin{array}{ccc}
\text{Rel}(\sigma) & \xrightarrow{i} & \text{P(Rel}(\sigma)) \\
\downarrow & & \downarrow \hat{\Pi} \\
\mathcal{M}_\sigma & \xrightarrow{i} & \text{P(M}_\sigma)
\end{array}$$
9. Spectrum Driven Clustering

We shall now make use of the abstract results of Section 8 to compute the globular clusters of a local convergent sequence.

In some sense globular clusters corresponds to the non-zero measure connected components at the limit. We shall see that we can track globular clusters and give an explicit formula for their limit size.

To achieve this, we shall first show that the moments of the distribution of the limit sizes of the globular clusters may be computed from Stone pairing, and then we shall deduce the distribution of the limit sizes of the globular clusters by standard Fourier analysis.

9.1. Spectrum. We start our analysis by the study of the limit sizes of the globular clusters.

Let \( \phi_d \) be the formula \( \text{dist}(x_1, x_2) \leq d \). Let \( \mathbf{A} \) be a local convergent sequence of \( \sigma \)-structures, and let \( D_{d,n} : A_n \rightarrow [0, 1] \) be the random variable

\[
D_{d,n}(v) = \langle \phi_d, A_n \rangle_v = v_{A_n}(N^d_{A_n}(v)).
\]

As obviously \( \phi_d \) implies \( \phi_{d+1} \) it follows from Lemma 28 that there exists random variables \( D_d \) and \( D \) such that \( D_{d,n} \overset{D}{\rightarrow} D_d \) and \( D \overset{D}{\rightarrow} D \) (which are limits in distribution, for \( n \rightarrow \infty \) and \( d \rightarrow \infty \), respectively).

Remark 6. The random variables \( D_{d,n} \) have here a concrete meaning, as the measure of the radius \( d \) ball centered at a random vertex. However, there is no particular meaning for the sample space of random variables \( D_d \) and \( D \) (as the existence of these were simply derived from convergence of characteristic functions).

Even if we intuitively interpret the random variables \( D_d \) and \( D \) as if they were built on a similar limit sample space, we have to take care in our argumentation that this interpretation is not \textit{a priori} justified.

We denote respectively by \( F_{d,n}, F_d \) and \( F \) the cumulative distribution functions of \( D_{d,n}, D_d \) and \( D \). According to Froda’s theorem, each \( F_d \) (and \( F \)) has at most countably many discontinuities. As \( D_d \overset{D}{\rightarrow} D \) (as \( d \rightarrow \infty \)) the functions \( F_d \) converge pointwise to \( F \) at every continuity point of \( F \). Similarly, for each \( d \in \mathbb{N} \), as \( D_{n,d} \overset{D}{\rightarrow} D_d \) (as \( n \rightarrow \infty \)) the functions \( F_{n,d} \) converge pointwise to \( F_d \) at every continuity point of \( F_d \).

We define \( \Lambda_d \) (resp. \( \Lambda \)) as the (at most countable) set of discontinuities of \( F_d \) (resp. \( F \)), and let \( \mathcal{R} = [0, 1] \setminus (\Lambda \cup \bigcup_{d \in \mathbb{N}} \Lambda_d) \). In other words, \( \mathcal{R} \) is the (counstable) set of points where all the considered limit cumulative functions (\( F \) and \( F_d \) for \( d \in \mathbb{N} \)) are continuous.

Remark 7. If \( d_1 < d_2 \) then for every integer \( n \) and ever real \( t \) it holds

\[
F_{d_1,n}(t) = \Pr(D_{d_1,n} \leq t) \geq \Pr(D_{d_2,n} \leq t) = F_{d_2,n}(t)
\]

and thus also \( F_{d_1} \geq F_{d_2} \geq F \).
We now take time for a useful simple lemma.

**Lemma 29.** Let \( t_1 < t_2 \) be in \( \mathcal{R} \), and let \( n \) and \( d_1 < d_2 \) be integers. Then

\[
F_{d_2,n}(t_2) - F_{d_1,n}(t_1) \leq \Pr(t_1 < D_{d_1,n} \leq D_{d_2,n} \leq t_2) \leq F_{d_1,n}(t_2) - F_{d_1,n}(t_1).
\]

Hence if \(|F(t_j) - F_{d_1}(t_j)| < \varepsilon\) and \(|F_{d_i}(t_j) - F_{d_i,n}(t_j)| < \varepsilon\) for \( i, j \in \{1, 2\} \) then

\[
|\Pr(t_1 < D_{d_1,n} \leq D_{d_2,n} \leq t_2) - (F(t_2) - F(t_1))| < 4\varepsilon.
\]

**Proof.** The right inequality is obvious as

\[
D(t) = \Pr(D_{d_1,n} \leq \lambda) \leq \Pr(D_{d_2,n} \leq \lambda) = D(t) + \Pr(D_{d_2,n} \leq \lambda, D_{d_1,n} > \lambda) = D(t) + \Pr(D_{d_1,n} < \lambda)
\]

due to the discontinuity points of \( D(t) \). Hence if \( \varepsilon \) would have a proper explicit limit structure, the random variable \( D \) would intuitively correspond to the measure of the connected component of a random element. Thus we expect \( D \) to be a discrete random variable, and that the probability that \( D = \lambda \) is the measure of the union of all connected components of measure \( \lambda \) hence an integral multiple of \( \lambda \). The aim of this part is to show that this intuitive notion of limit connected components is captured by the concept of globular clusters. This setting will not only ground the above intuition, but will also allow to track the formation of the limit connected components down to the structures in the sequence.

Hence our first step is to prove that \( D \) is a purely discrete random variable, that is that its cumulative distribution function \( F \) is constant except at its (at most countably many) discontinuity points. This we shall do now.

**Lemma 30.** The spectrum distribution of a local-convergent sequence of finite structures is discrete and its associated mass probability function \( p : [0, 1] \rightarrow [0, 1] \) defined by

\[
p(x) = F(x) - \lim_{\varepsilon \to 0} F(x - \varepsilon).
\]

is such that that for every \( x \in [0, 1] \), either \( p(x) = 0 \) or \( p(x) \geq x \).

**Proof.** We shall prove that for \( t_1 < t_2 \) in \( \mathcal{R} \), either \( F(t_1) = F(t_2) \) or \( F(t_2) - F(t_1) \geq t_1 \). It will follow, by cutting the interval \([t_1, t_2]\) recursively, that \( F \) is constant except at its discontinuity points, and that the mass probability function \( p \) satisfies \( p(x) \geq x \) at every point \( x \) where \( p(x) \neq 0 \).
So let $t_1 < t_2$ be in $\mathcal{R}$ and such that $F(t_1) < F(t_2)$, and let $0 < \epsilon < (F(t_2) - F(t_1))/4$. As $D \xrightarrow{d} D$ (as $d \to \infty$) there exists $d$ such that $|F_d(t_1) - F(t_1)| < \epsilon$ and $|F_d(t_2) - F(t_2)| < \epsilon$. Moreover, as $D_{n,d} \xrightarrow{d} D_d$ (for fixed $d$ and $n \to \infty$) there exists $n$ such that $|F_{d,n}(t_1) - F_d(t_1)| < \epsilon$, $|F_{d,n}(t_2) - F_d(t_2)| < \epsilon$, $|F_{2d,n}(t_1) - F_{2d}(t_1)| < \epsilon$ and $|F_{2d,n}(t_2) - F_{2d}(t_2)| < \epsilon$.

According to Lemma 29 it holds

$$\Pr(t_1 < D_{d,n} \leq D_{2d,n} \leq t_2) - (F(t_2) - F(t_1)) < 4\epsilon$$

thus $\Pr(t_1 < D_{d,n} \leq D_{2d,n} \leq t_2) > 0$. Hence there exists $\nu \in A_n$ such that $t_1 < D_{d,n}(\nu) \leq D_{2d,n}(\nu) \leq t_2$. For every $x \in N^d_{A_n}(\nu)$ it holds $N^d_{A_n}(x) \subseteq N^{2d}_{A_n}(\nu)$ hence $D_{d,n}(x) \leq D_{2d,n}(\nu) \leq t_2$. Also, $N^{2d}_{A_n}(\nu) \supseteq N^d_{A_n}(\nu)$ thus $D_{2d,n}(x) \geq D_{d,n}(\nu) \geq t_1$. As this holds for every $x \in N^d_{A_n}(\nu)$, we get $N^d_{A_n}(\nu) \subseteq \{x : t_1 < D_{2d,n}(x) \text{ and } D_{d,n}(x) \leq t_2\}$. Thus we have

$$\nu_{A_n}(N^d_{A_n}(\nu)) - \Pr(t_1 < D_{2d,n} \leq t_2) \leq \Pr(t_1 < D_{2d,n} \text{ and } D_{d,n} \leq t_2) - \Pr(t_1 < D_{2d,n} \leq t_2)$$

$$\leq \Pr(D_{d,n} \leq t_2) - \Pr(D_{2d,n} \leq t_2)$$

$$= F_{d,n}(t_2) - F_{2d,n}(t_2)$$

$$< 4\epsilon.$$

Hence $\Pr(t_1 < D_{2d,n} \leq t_2) > \nu_{A_n}(N^d_{A_n}(\nu)) - 4\epsilon > t_1 - 4\epsilon$. Hence $F(t_2) - F(t_1) > t_1 - 8\epsilon$. By letting $\epsilon \to 0$, we get $F(t_2) - F(t_1) \geq t_1$ as claimed. \qed

Recall that $\Lambda$ is the set of discontinuities of $F$, that is the set of $x \in [0, 1]$ such that $p(x) \neq 0$. Note that it follows from Lemma 30 that for every integer $z$ there exists at most $z$ values $\lambda \in \Lambda$ with $\lambda \geq 1/z$.

The next lemma will ground our intuition that $p(\lambda)$ should be an integral multiple of $\lambda$. Indeed, we will prove later that $p(\lambda)/\lambda$ is the number of disjoint globular clusters with limit measure $\lambda$.

**Lemma 31.** Let $\lambda \in \Lambda$. Then $p(\lambda)/\lambda \in \mathbb{N}$.

**Proof.** Let $0 < \epsilon < \lambda^2/11$. Fix $t_1, t_2 \in \mathcal{R}$ with $0 < t_1 < \lambda < t_2$, $t_2 - t_1 < \epsilon$, and such that $\lambda$ is the only discontinuity point of $F$ on $[t_1, t_2]$ (hence $p(\lambda) = F(t_2) - F(t_1)$).

Then there exist $\delta = \delta(\epsilon, t_1, t_2)$ such that for every $d \geq \delta$ it holds $|F_{t_1} - F_{kd,t_1}| < \epsilon$ for every $1 \leq k \leq 4$ and every $i \in \{1, 2\}$, and there exists $\eta = \eta(\epsilon, t_1, t_2, d)$ such that for every $n \geq \eta$ it holds $|F_{kd,n}(t_i) - F_{kd}(t_i)| < \epsilon$ for every $1 \leq k \leq 4$ and every $i \in \{1, 2\}$.

We prove by contradiction that no two vertices $v_1, v_2 \in A_n$ exist such that

$$t_1 < D_{d,n}(v_1) \leq D_{4d,n}(v_1) \leq t_2,$$

$$t_1 < D_{d,n}(v_2) \leq D_{4d,n}(v_2) \leq t_2,$$

$$2d < \text{dist}(v_1, v_2) \leq 3d.$$
Assume the contrary. Then $N_{A_n}^{4d}(v_1)$ contains the disjoint union of $N_{A_n}^{d}(v_1)$ and $N_{A_n}^{d}(v_2)$ thus

$$v_{A_n}(N_{A_n}^{4d}(v_1)) > 2t_1 > t_1 + \lambda - \epsilon = (t_1 + \epsilon) + (\lambda - 2\epsilon) > t_2,$$

contradicting $D_{4d,n}(v_1) \leq t_2$.

Let $S = S_{t_1,t_2,d}(n)$ be a maximal set of vertices $v \in A_n$, pairwise at distance greater than $3d$, and such that

$$t_1 < D_{d,n}(v) \leq D_{4d,n}(v) \leq t_2.$$

First note that for $v, v' \in S$ the balls $N_{A_n}^{d}(v)$ and $N_{A_n}^{d}(v')$ do not intersect hence

$$1 \geq v_{A_n}(\bigcup_{v \in S} N_{A_n}^{d}(v)) = \sum_{v \in S} D_{d,n}(v) > t_1 |S| > (\lambda - \epsilon)|S|.$$

Thus $|S| < 1/(\lambda - \epsilon)$.

Also every vertex $w$ such that $t_1 < D_{d,n}(w) \leq D_{4d,n}(w) \leq t_2$ belongs to $\bigcup_{v \in S} N_{A_n}^{2d}(v) = N_{A_n}^{2d}(S)$. It follows that $\Pr(t_1 < D_{d,n} \leq D_{4d,n} \leq t_2) \leq t_2 |S|$. Also,

$$\Pr(t_1 < D_{2d,n} \leq t_2) \geq v_{A_n}(N_{A_n}^{2d}(S)) = \sum_{s \in S} D_{d,n}(s) > t_1 |S| > \lambda |S| - \epsilon/(\lambda - \epsilon).$$

As

$$\Pr(t_1 < D_{2d,n} \leq t_2) \leq \Pr(t_1 < D_{d,n} \leq D_{4d,n} \leq t_2) \leq t_2 |S| < \lambda |S| + \epsilon/(\lambda - \epsilon),$$

we get

$$|\Pr(t_1 < D_{2d,n} \leq t_2) - \lambda |S|| < \epsilon/(\lambda - \epsilon).$$

As

$$|\Pr(t_1 < D_{2d,n} \leq t_2) - p(\lambda)| < 4\epsilon$$

we deduce

$$|p(\lambda) - \lambda |S|| < (4 + 1/(\lambda - \epsilon))\epsilon.$$

As $\epsilon < \lambda^2/11$, it holds $|p(\lambda) - |S|\lambda| < \lambda/2$, thus $|S| = |S_{t_1,t_2,d}(n)|$ is constant for all the values $t_1, t_2, d, n$ consistent with $0 < \epsilon < \lambda^2/11$. Denoting $m(\lambda)$ this common value of $|S_{t_1,t_2,d}(n)|$, and by letting $\epsilon \to 0$, we get $p(\lambda) = m(\lambda)\lambda$ thus $p(\lambda)/\lambda \in \mathbb{N}$. 

We now define several functions, which will be of key importance in our precise definition and analysis of the globular clusters.

Let us fix $\lambda \in \Lambda$.

**Definition of $\epsilon_z$.** For $z \in \mathbb{N}$, we define

$$\epsilon_z = 2^{-z}.$$

**Definition of $z_0(\lambda)$.** We define

$$z_0(\lambda) = \lfloor 5 - 2 \log_2 \lambda \rfloor.$$

(Thus $\epsilon_{z_0(\lambda)} \leq \lambda^2/32$.)
Definition of $\alpha_z(\lambda)$ and $\beta_z(\lambda)$. We define

$$\alpha_1(\lambda) < \alpha_2(\lambda) < \ldots < \lambda < \ldots < \beta_2(\lambda) < \beta_1(\lambda),$$

such that $\Lambda \cap [\alpha_1(\lambda), \beta_1(\lambda)] = \{\lambda\}$, every $\alpha_z(\lambda)$ and $\beta_z(\lambda)$ belong to $\mathcal{R}$, and such that for every $z \in \mathbb{N}$ it holds

$$(8) \quad |\beta_z(\lambda) - \alpha_z(\lambda)| < \varepsilon_z.$$  

Definition of $\delta_z(\lambda)$. As $D_d \overset{D}{\to} D$ (as $d \to \infty$) we can define integers $\delta_1(\lambda) < \delta_2(\lambda) < \ldots$ such that for every $z \in \mathbb{N}$ and every $d \geq \delta_z(\lambda)$ it holds

$$(9) \quad |F_d(\alpha_z(\lambda)) - F(\alpha_z(\lambda))| < \varepsilon_z$$

$$(10) \quad |F_d(\beta_z(\lambda)) - F(\beta_z(\lambda))| < \varepsilon_z$$

Definition of $\eta_z(\lambda)$. As $D_{n,d} \overset{D}{\to} D_d$ (for fixed $d$ and as $n \to \infty$) we can define integers $\eta_1(\lambda) < \eta_2(\lambda) < \ldots$ such that for every $z \in \mathbb{N}$, every $n \geq \eta_z(\lambda)$ and every integer $k \in \{1, \ldots, 8\}$ it holds

$$(11) \quad |F_{k\delta_z(\lambda),n}(\alpha_z(\lambda)) - F_{k\delta_z(\lambda)}(\alpha_z(\lambda))| < \varepsilon_z$$

$$(12) \quad |F_{k\delta_z(\lambda),n}(\beta_z(\lambda)) - F_{k\delta_z(\lambda)}(\beta_z(\lambda))| < \varepsilon_z$$

We now define some sequences of sets. The sets $Z^\lambda_n$ will anticipate our construction of globular clusters, by giving a rough approximate of them. Then the set $S^\lambda_n$ will collect a “center” for each of the “component” of size $\lambda$.

Definition of $Z^\lambda_n$. For $n, z \in \mathbb{N}$ we define subset $Z^\lambda_n$ as follows:

- If $n < \eta_z$ then $Z^\lambda_n = \emptyset$;
- Otherwise, $Z^\lambda_n$ is the set of all elements of $A_n$ such that

$$(13) \quad D_{8\delta_z,n}(v) \leq \beta_z(\lambda)$$

$$(14) \quad D_{\delta_z,n}(v) > \alpha_z(\lambda) \quad (\forall z' \in \{z_0(\lambda), \ldots, z\})$$

Definition of $S^\lambda_n$. We define $S^\lambda_n$ as a maximal set of vertices $v \in Z^\lambda_n$, pairwise at distance at least $7\delta_z$, where $z$ is (implicitly) defined by $\eta_z \leq n < \eta_{z+1}$.

We take time for few remarks:

Remark 8. Note that (13) implies $D_{8\delta_z,n}(v) \leq \beta_{z'}(\lambda)$ for every $1 \leq z' \leq z$. Also (14) becomes clearly more and more restrictive as $z$ grows. Hence for every $z \geq z_0(\lambda)$ and every $n \in \mathbb{N}$ such that $\eta_z \leq n < \eta_{z+1}$ it holds

$$(15) \quad Z^\lambda_{n,z_0(\lambda)} \supset Z^\lambda_{n,z_0(\lambda)+1} \supset \ldots \supset Z^\lambda_n \supset \emptyset \quad (\forall z \in \{z_0(\lambda), \ldots, z\})$$

Remark 9. According to the definitions of $\delta_z$ and $\eta_z$, it holds

$$|F_{8\delta_z,n}(\beta_z(\lambda)) - F(\beta_z(\lambda))| < 2\varepsilon_z$$
Remark 10. If \( z' < z'' \) then, according to Lemma 29 it holds
\[
\Pr(\alpha_{z'}(\lambda) < D_{\delta_{z'}(\lambda),n} \leq D_{\delta_{z''}(\lambda),n} \leq \alpha_{z''}(\lambda)) < 4e_{z'}
\]
Thus
\[
\Pr(D_{\delta_{z}(\lambda),n} > \alpha_{z}(\lambda)) - \Pr\left( \bigwedge_{z'<z_0} D_{\delta_{z'}(\lambda),n} > \alpha_{z'}(\lambda) \right) < 4 \sum_{z'=z_0}^{z} e_{z'} = 2^{2-z_0}.
\]
It follows that
\[
\nu_{\Lambda_n}(Z^{\lambda,z}_n) \geq p(\lambda) - 4e_z - 2^{2-z_0}.
\]

We now prove that, as wanted, the number of elements of \( S^n_{\lambda} \) is (for sufficiently large \( \lambda \)) the anticipated number of globular clusters of size \( \lambda \).

**Lemma 32.** For every \( \lambda \in \mathcal{R} \) and every \( n \geq \eta_{\lfloor \lambda^{-1} \rfloor} \) it holds \( |S^n_{\lambda}| = p(\lambda)/\lambda \).

**Proof.** Note that obviously, as \( \nu_{\Lambda_n}(N^{\delta_{z}(\lambda)}_{\Lambda_n}(s)) < \lambda + \varepsilon_z \) holds for every \( s \in S^n_{\lambda} \), we get
\[
|S^n_{\lambda}| \geq \frac{|Z^{\lambda,z}_n|}{\lambda + \varepsilon_z} \geq \frac{p(\lambda) - 2^2z - 2^2z_0(\lambda)}{\lambda + 2^2z} = \frac{p(\lambda)}{\lambda} - \frac{2^2z - 2^2z_0(\lambda)}{\lambda + 2^2z} > \frac{p(\lambda)}{\lambda} - 1,
\]
hence \( |S^n_{\lambda}| \geq p(\lambda)/\lambda \). On the other hand, for every \( s \in S^n_{\lambda} \) it holds \( D_{\delta_z(\lambda),n}(s) > \alpha_z(\lambda) \) and \( D_{\delta_z(\lambda),n}(s) \leq \beta_z(\lambda) \) thus for every \( v \in N^{\delta_z(\lambda)}_{\Lambda_n}(S^n_{\lambda}) \) it holds \( \alpha_z(\lambda) < D_{\delta_z(\lambda),n}(s) \leq \beta_z(\lambda) \) thus
\[
(\lambda - \varepsilon_z)|S^n_{\lambda}| < \nu_{\Lambda_n}(N^{\delta_z(\lambda)}_{\Lambda_n}(S^n_{\lambda})) \leq \Pr(\alpha_z(\lambda) < D_{\delta_z(\lambda),n}(s) \leq \beta_z(\lambda)) < p(\lambda) + 4\varepsilon_z
\]
hence
\[
|S^n_{\lambda}| < \frac{p(\lambda)}{\lambda} + 1.
\]
Altogether, it follows that \( |S^n_{\lambda}| = p(\lambda)/\lambda \). \( \square \)

We are now ready to define sets gathering all the “components” with limit measure \( \lambda \). We will prove that they define universal clusters.

**Definition of \( C^\lambda_n \).** For \( \lambda \in \Lambda \) and \( n \in \mathbb{N} \) we define
\[
C^\lambda_n = \begin{cases} 
\emptyset, & \text{if } n < \eta_{z_0(\lambda)} \\
N^{2\delta_z}_{\Lambda_n}(S^n_{\lambda}), & \text{otherwise, if } z \text{ is such that } \eta_z \leq n < \eta_{z+1}
\end{cases}
\]

The sets \( C^\lambda_n \) will be the building block for the construction of our clusters. Lemmas 33 to 38 will be used to prove that the sequences \( C^\lambda_n \) define a clustering of \( \mathbf{A} \) into countably many universal clusters plus a residual cluster. The general aspects of the sets \( Z^{\lambda,z}_n, S^n_{\lambda}, \) and \( C^\lambda_n \) we tried to visualize by Fig. 7.

Similarly to Lemma 31 we prove...
Lemma 33. Let \( \lambda \in \Lambda \), let \( z \geq z_0(\lambda) \), and let \( \eta_z \leq n < \eta_{z+1} \).

Then \( Z_{n}^{\lambda,z} \subseteq C_{n}^{\lambda} \).

Proof. Assume for contradiction that there exists an element \( v \in Z_n^{\lambda,z} \setminus C_n^{\lambda} \). By the maximality of \( S_n^{\lambda} \), we get that \( v \) is at distance at most \( 7\delta_z \) from some element \( u \in S_n^{\lambda} \). Moreover, \( \text{dist}(u, v) > 2\delta_z \) as \( v \not\in N_{A_n}^{2\delta_z}(S_n^{\lambda}) \). Then \( N_{A_n}^{8\delta_z}(v) \) contains the disjoint union of \( N_{A_n}^{\delta_z}(v) \) and \( N_{A_n}^{\delta_z}(u) \) thus

\[
\nu_{A_n}(N_{A_n}^{8\delta_z}(v)) > 2\alpha_z(\lambda) > \alpha_z(\lambda) + \lambda - e_2 = (\alpha_z(\lambda) + e_z) + (\lambda - 2e_z) > \beta_z(\lambda),
\]

contradicting \( D_{8\delta_z}(v) \leq \beta_z(\lambda) \).

We now prove that our sets \( C_n^{\lambda} \) are pairwise disjoint.

Lemma 34. Let \( \lambda < \lambda' \) be two elements of \( \Lambda \), let \( z \geq \max(z_0(\lambda), [1 - \log_2(\lambda' - \lambda)]) \), and let \( \eta_z \leq n < \eta_{z+1} \).

Then \( C_n^{\lambda} \cap C_n^{\lambda'} = \emptyset \).

Proof. Assume for contradiction that there exists an element \( v \in C_n^{\lambda} \cap C_n^{\lambda'} \). Then there exists \( u \in S_n^{\lambda} \) and \( u' \in S_n^{\lambda'} \) such that \( v \in N_{A_n}^{2\delta_z}(u) \cap N_{A_n}^{2\delta_z}(u') \) hence \( \text{dist}(u, u') \leq 4\delta_z \). It follows that \( N_{A_n}^{8\delta_z}(u') \subseteq N_{A_n}^{8\delta_z}(u) \) hence \( \alpha_z(\lambda') \leq \beta_z(\lambda) \) thus \( |\lambda - \lambda'| < 2.2^{-z} \), contradicting our choice of \( z \).
We now prove that the measure of $C_n^\lambda$ is concentrated around the limit measure $p(\lambda)$.

**Lemma 35.** Let $\lambda \in \Lambda$, let $z \geq z_0(\lambda)$, and let $\eta_z \leq n < \eta_{z+1}$.

Then $|v_{A_n}(C_n^\lambda) - p(\lambda)| < 2^{-z} p(\lambda)/\lambda$.

**Proof.** Note that

\[ v_{A_n}(C_n^\lambda) = \sum_{s \in S_n^\lambda} v_{A_n}(N_{A_n}^{2\delta_z}(s)). \]

Hence, as $|S_n^\lambda| = p(\lambda)/\lambda$ it holds

\[ |v_{A_n}(C_n^\lambda) - p(\lambda)| < 2^{-z} p(\lambda)/\lambda. \]

The next lemma not only shows that the outer boundary of $C_n^\lambda$ is negligible (what is required in order for $C_n^\lambda$ to be a cluster) but also that the neighborhood of these outer boundaries are so small that their sum will also be small (what we will make use of in lemma 37).

**Lemma 36.** Let $\lambda \in \Lambda$, let $n \geq \lceil \eta_{\lceil \lambda \rceil - 1} \rceil$, and let $z \in \mathbb{N}$ be such that $\eta_z \leq n < \eta_{z+1}$. Then it holds

\[ v_{A_n}(N_{A_n}^{\delta_z}(\partial_{A_n} C_n^\lambda)) < 2^{1-z} p(\lambda)/\lambda. \]

In particular, $\partial_{A} C_n^\lambda \approx 0$.

**Proof.** As elements of $S_n^\lambda$ are pairwise at distance at least $7\delta_z$, it holds

\[ N_{A_n}^{\delta_z}(\partial_{A_n} C_n^\lambda) = \bigcup_{v \in S_n^\lambda} (N_{A_n}^{3\delta_z}(v) \setminus N_{A_n}^{\delta_z}(v)) \]

(where $\cup$ denotes a disjoint union). As $v \in S_n^\lambda$ it holds

\[ v_{A_n}(N_{A_n}^{\delta_z}(v)) = D_{\delta_z,n}(v) > \alpha_z(\lambda) \]

\[ v_{A_n}(N_{A_n}^{3\delta_z}(v)) \leq D_{8\delta_z,n}(v) \leq \beta_z(\lambda) \]

Hence

\[ v_{A_n}(N_{A_n}^{3\delta_z}(v) \setminus N_{A_n}^{\delta_z}(v)) < \epsilon_z \]

Thus

\[ v_{A_n}(N_{A_n}^{\delta_z}(\partial_{A_n} C_n^\lambda)) < |S_n^\lambda|\epsilon_z < 2\epsilon_z p(\lambda)/\lambda \]

**Lemma 37.** Let $n \in \mathbb{N}$, let $\lambda \in \Lambda$ be minimum such that $n \geq \eta_{z_0(\lambda)}$. Let $z$ be defined by $\eta_z \leq n < \eta_{z+1}$, and let

\[ W_n = \{v : D_{\delta_z,n}(v) > \alpha_z(\lambda)\} \setminus \bigcup_{\lambda \in \Lambda} C_n^\lambda. \]

Then

\[ v_{A_n}(N_{A_n}^{\delta_z}(W_n)) \leq 2^{-z}(1 + 3/\lambda). \]

In particular, $W \supseteq \partial_{A}(\bigcup_{\lambda \in \Lambda} C_n^\lambda)$ and $W \approx 0$. 
Proof. Let \( F_n = \{ v : D_{\delta z,n}(v) > \alpha_z(\lambda) \} \). Then \( N_{\lambda_n}^{\delta z}(F_n) \subseteq \{ v : D_{\delta z,n}(v) > \alpha_z(\lambda) \} \). Hence \( v_{A_n}(N_{\lambda_n}^{\delta z}(F_n)) \leq 1 - F_{2\delta z,n}(\alpha_z(\lambda)) \). It follows that
\[
\begin{align*}
\nu_{A_n}(N_{\lambda_n}^{\delta z}(W_n)) & \leq \nu_{A_n}(N_{\lambda_n}^{\delta z}(F_n)) - \sum_{\lambda' \geq \lambda} \nu_{A_n}(C_{\lambda'}^n) + \sum_{\lambda' \geq \lambda} \nu_{A_n}(N_{\lambda_n}^{\delta z}(\partial_{\lambda'} A_n C_{\lambda'}^n)) \\
& \leq \epsilon_z + \frac{2^{-z}}{\lambda} + \frac{2^{1-z}}{\lambda}.
\end{align*}
\]

We now are ready for our last lemma needed to prove that the sequences \( C^\lambda \) define a clustering of \( A \) into countably many universal clusters plus a residual cluster.

**Lemma 38.** For each \( \lambda \in \Lambda \) the sequence \( C^\lambda = (C_n^\lambda)_{n \in \mathbb{N}} \) is a cluster.

Proof. Let \( \phi \) be an \( r \)-local strongly local formula with free variables \( x_1, \ldots, x_q \). For \( d \in \mathbb{N} \) let \( \Psi_d \) be the following formula with \( q + 1 \) free variables
\[
\Psi_d : \quad \phi(x_2, \ldots, x_{q+1}) \land \bigwedge_{i=2}^{q+1} \text{dist}(x_1, x_i) \leq d.
\]

Note that if \( d_1 < d_2 \) and \( v \in A_n \) then
\[
0 \leq \langle \Psi_{d_2}, A_n \rangle_v - \langle \Psi_{d_1}, A_n \rangle_v \leq q \langle D_{d_2,n}(v) - D_{d_1,n}(v) \rangle.
\]

For \( \lambda \in \Lambda \) we consider an integer \( z_1 \) such that \( \delta_{z_1} > r \) and \( z_1 \geq z_0(\lambda) \), an integer \( z \geq z_1 \) and \( \eta_z \leq n < \eta_{z+1} \). Then the following holds: for every \( s \in S_n^\lambda \) and every \( x \in N_{\lambda_n}^{\delta z_1}(s) \), it holds \( N_{\lambda_n}^d(x) \subseteq N_{\lambda_n}^{d+\delta z_1}(s) \) and \( N_{\lambda_n}^d(s) \subseteq N_{\lambda_n}^{d+\delta z_1}(x) \) we get
\[
\langle \Psi_{d-\delta z_1}, A_n \rangle_s \leq \langle \Psi_{d}, A_n \rangle_x \quad \text{and} \quad \langle \Psi_{d}, A_n \rangle_x \leq \langle \Psi_{d+\delta z_1}, A_n \rangle_s.
\]

It follows that
\[
\begin{align*}
\langle \Psi_{2\delta z_1}, A_n \rangle_x - \langle \Psi_{2\delta z_1}, A_n \rangle_s & \leq \langle \Psi_{3\delta z_1}, A_n \rangle_s - \langle \Psi_{2\delta z_1}, A_n \rangle_s \\
& \leq q \Pr(2\delta z_1 \leq \text{dist}(x, s) \leq 3\delta z_1) \\
& \leq q \langle D_{8\delta z_1,n}(s) - D_{\delta z_1,n}(s) \rangle \\
& < q(\beta_z(\lambda) - \alpha_z(\lambda)) \\
& < q(\epsilon_{z_1} + \epsilon_z)
\end{align*}
\]

and
\[
\begin{align*}
\langle \Psi_{2\delta z_1}, A_n \rangle_s - \langle \Psi_{2\delta z_1}, A_n \rangle_x & \leq \langle \Psi_{2\delta z_1}, A_n \rangle_s - \langle \Psi_{\delta z_1}, A_n \rangle_s \\
& \leq q \Pr(\delta z_1 \leq \text{dist}(x, s) \leq 2\delta z_1) \\
& < q(\epsilon_{z_1} + \epsilon_z).
\end{align*}
\]

Thus
\[
|\langle \Psi_{2\delta z_1}, A_n \rangle_x - \langle \Psi_{2\delta z_1}, A_n \rangle_s| < q(\epsilon_{z_1} + \epsilon_z).
\]
Also,
\[
|\langle \Psi_{2 \delta_z}, A_n \rangle_s - \langle \Psi_{2 \delta_{z_1}}, A_n \rangle_s| \leq q(D_{\delta_z,n}(s) - D_{\delta_{z_1},n}(s)) \\
\leq q(D_{8 \delta_z,n}(s) - D_{\delta_{z_1},n}(s)) \\
\leq q(\beta_z(\lambda) - \alpha_{z_1}(\lambda)) \\
< q(\epsilon_{z_1} + \epsilon_z).
\]

Moreover,
\[
|\langle \Psi_{2 \delta_z}, A_n \rangle_s - \langle \Psi_{2 \delta_z}, A_n - \partial_{A_n} C_n^\lambda \rangle_s| < 4q\epsilon_x p(\lambda)/\lambda.
\]

and
\[
\langle \phi, A_n [C_n^\lambda] \rangle = \sum_{s \in S_n^i} \langle \Psi_{2 \delta_z}, A_n - \partial_{A_n} C_n^\lambda \rangle_s / \nu_{A_n}(C_n^\lambda)^p.
\]

Thus, as
\[
|\nu_{A_n}(C_n^\lambda) - \lambda| < \epsilon_x p(\lambda)/\lambda,
\]
it holds
\[
\mathbb{E} [\langle \Psi_{2 \delta_{z_1}}, A_n \rangle_v 1_{Z_n^{\lambda,z_1}}(v)] = \frac{\sum_{v \in Z_n^{\lambda,z_1}} \nu_{A_n}(v) \langle \Psi_{2 \delta_z}, A_n \rangle_v}{\nu_{A_n}(Z_n^{\lambda,z_1})} \\
\approx \frac{1}{\lambda} \sum_{v \in C_n^\lambda} \langle \Psi_{2 \delta_z}, A_n \rangle_v \\
\approx \sum_{s \in S_n^i} \langle \Psi_{2 \delta_z}, A_n \rangle_s \\
\approx \lambda^p \langle \phi, A_n [C_n^\lambda] \rangle
\]

Let \( H_{z_1,n} \) be the (multivariate) cumulative distribution function of
\[
(\langle \Psi_{\delta_{z_1}}, A_n \rangle_v, 1 - D_{\delta_{z_0}(\lambda)}, \ldots, 1 - D_{\delta_{z_1}}, D_{8 \delta_{z_1}}).
\]

According to its definition we have \( v \in Z_n^{\lambda,z_1} \) if and only if
\[
(1 - D_{\delta_{z_0}(\lambda)}, \ldots, 1 - D_{\delta_{z_1}}, D_{8 \delta_{z_1}}) \in [0, 1 - \alpha_{z_0}(\lambda)] \times \cdots \times [0, 1 - \alpha_{z_1}(\lambda)] \times [0, \beta_{z_1}(\lambda)].
\]

Thus
\[
\Pr[\langle \Psi_{\delta_{z_1}}, A_n \rangle_v \leq x \text{ and } v \in Z_n^{\lambda,z_1}] = H_{n,z_1}(x, 1 - \alpha_{z_0}(\lambda), \ldots, 1 - \alpha_{z_1}(\lambda), \beta_{z_1}(\lambda)).
\]

It follows that
\[
\mathbb{E} [\langle \Psi_{2 \delta_{z_1}}, A_n \rangle_v 1_{Z_n^{\lambda,z_1}}(v)] = \int_0^1 \Pr[\langle \Psi_{\delta_{z_1}}, A_n \rangle_v \leq x \text{ and } v \in Z_n^{\lambda,z_1}] \, dx \\
= \int_0^1 1 - H_{n,z_1}(x, 1 - \alpha_{z_0}(\lambda), \ldots, 1 - \alpha_{z_1}(\lambda), \beta_{z_1}(\lambda)) \, dx.
\]
According to Lemma 27 there exists a (vector) random variable $V_{z_1}$ such that

$$(\langle \Psi_{\delta_{z_1}}, A_n \rangle, 1 - D_{\delta_{z_0}(\lambda)}, \ldots, 1 - D_{\delta_{z_1}}, D_{\delta_{z_1}}) \overset{D}{\to} V_{z_1}.$$ 

Let $H$ be the cumulative distribution function of $V_{z_1}$. Then, as $n \to \infty$ it holds

$$\lim_{n \to \infty} \mathbb{E}[\langle \Psi_{2\delta_{z_1}}, A_n \rangle, 1_{Z_{n}^{z_{1}}}(\nu)] = \int_{0}^{1} 1 - H(x, 1 - \alpha_{z_0(\lambda)}(\lambda), \ldots, 1 - \alpha_{z_1}(\lambda)) \, dx.$$ 

As $|\mathbb{E}[\langle \Psi_{2\delta_{z_1}}, A_n \rangle, 1_{Z_{n}^{z_{1}}}(\nu)] - \lambda^p(\phi, A_n[C_n])|$ goes to 0 when $z_1$ goes to infinity (and $n$ grows in consequence), we get that $\langle \phi, A_n[C_n] \rangle$ converges hence $C_i$ is a cluster. □

We are now ready to prove our first clustering result:

**Lemma 39.** Let $A$ be a local convergent sequence of $\sigma$-structures. Let $\sigma^+$ be the signature obtained from $\sigma$ by the addition of countably many unary symbols $M_R$ and $M_i$ ($i \in \mathbb{N}$). Then marking by $M_i$ the cluster $C_n^{\lambda_i}$ (where $\lambda_1 > \lambda_2 > \ldots$ are the elements of $\Lambda$ order in decreasing order) and by $M_0$ the sequence of sets

$$R = A \setminus W \setminus \bigcup_{\lambda \in \Lambda} C^\lambda_i$$

we obtain clustering $L(A)$ of $A$ with the following properties:

- For every $i \in \mathbb{N}$, $(M_i(L(A_n)))_{n \in \mathbb{N}}$ is a universal globular cluster, and $M_i(L(A_n))$ asymptotically consists in a set inducing $p(\lambda_i)/\lambda_i$ disjoint connected substructures, each of measure $\lambda_i + o(1)$ in $A_n$.
- $(M_R(A_n^+))_{n \in \mathbb{N}}$ is a residual cluster.

**Proof.** That $L(A)$ is clustering follows from Lemma 22. That $C_i$ is a universal cluster is trivial as the constructions and proofs can be achieved the same way (with same result) in any conservative lift of $A$. The sequence $R$ is obviously residual. □

We are now ready to prove Theorem 1, which we state now in the following more precise form.
Theorem 7. Let $\mathbf{A}$ be a local convergent sequence of $\sigma$-structures. Then there exists a signature $\sigma^+$ (obtained from $\sigma$ by the addition of countably many unary symbols $M_{i,j,k}$ ($i \in \mathbb{N}$, $1 \leq j \leq a_i$, $1 \leq k \leq b_{i,j}$), of a unary symbol $M_R$ and of unary symbol $M_S$), a sequence $\lambda_1 > \lambda_2 > \ldots$ a positive reals and a clustering $L(\mathbf{A})$ of $\mathbf{A}$ with the following properties:

- For every $i \in \mathbb{N}$, $1 \leq j \leq a_i$, and $1 \leq k \leq b_{i,j}$, $G_{i,j,k} = M_{i,j,k}(L(\mathbf{A}))$ is a globular cluster of $\mathbf{A}$ such that $\lim_{n \to \infty} v_\mathbf{A}(G_{i,j,k}^d) = \lambda_i$, that is a cluster such that for every positive real $\epsilon$ there is an integer $d$ which satisfies
  \[
  \lambda_i - \epsilon < \liminf_{n \to \infty} \max_{v_n \in G_{i,j,k}^d} v_\mathbf{A}(N_{i,j,k}^d) \leq \lim_{n \to \infty} v_\mathbf{A}(G_{i,j,k}^d) = \lambda_i.
  \]

- $R = M_R(L(\mathbf{A}))$ is a residual cluster of $\mathbf{A}$, that is a cluster such that for every integer $d$ it holds
  \[
  \limsup_{n \to \infty} \max_{v_n \in G_{i,j,k}^d} v_\mathbf{A}(N_{i,j,k}^d) = 0.
  \]

- The sequence $S$ is negligible, that is such that for every integer $d$ it holds
  \[
  \limsup_{n \to \infty} v_\mathbf{A}(N_{i,j,k}^d) = 0.
  \]

- The marks partition the sets $A_n$ is a stable way, that is
  \[
  \lim_{n \to \infty} v_\mathbf{A}(R) + \sum_{i \geq 1} \lim_{n \to \infty} v_\mathbf{A}(G_{i,j,k}^d) = 1.
  \]

- Clusters $G_{i,j,k}$ and $G_{i',j',k'}$ are interweaving (i.e. $G_{i,j,k} \nparallel G_{i',j',k'}$) if and only if $i = i'$ and $j = j'$.

- The clusters $\bigcup_{k=1}^{b_{i,j}} G_{i,j,k}$ (grouping interweaving clusters) are universal.

- The number $N_i = \sum_{j=1}^{a_i} b_{i,j}$ of clusters with limit measure $\lambda_i$ is
  \[
  N_i = \frac{1}{\lambda_i} \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{+T} \left[ \sum_{w \geq 1} \left( \lim_{d \to \infty} \int_{S_\sigma} k(\psi_{d,w}) d\mu \right) (is)^w \right] \frac{1}{w!} e^{-is} ds,
  \]
  where $\psi_{d,w}$ is the formula
  \[
  \psi_{d,w}(x_1, \ldots, x_{w+1}) := \bigwedge_{i=1}^w \text{dist}(x_1, x_i) \leq d.
  \]

Proof. By construction, the number of connected components of $A_n[C_n^d]$ is asymptotically $p(\lambda)/\lambda$ and each of these connected components has asymptotically measure $\lambda$. Let $B_{n,1}, \ldots, B_{n,k_n}$ be the connected components of $A_n[C_n^d]$. If there is a local formula $\phi$ such that
  \[
  \lim_{n \to \infty} \min_{i} \langle \phi, B_i \rangle \neq \lim_{n \to \infty} \min_{i} \langle \phi, B_i \rangle
  \]
we can break $C^\lambda$ into smaller universal clusters. At the end of the day, we get a clustering of $A$ into countably many clusters, such that each cluster $C^i$ has asymptotically $k_i$ connected components with same asymptotic measure and same asymptotic profile. It follows that $C^i$ is the disjoint union of $k_i$ interweaving clusters.

The statement giving the number $N_i$ of clusters with measure $\lambda_i$ is due to the equality $N_i = p(\lambda_i)\lambda_i$ and the application of Lévy’s theorem (Theorem 5) for the computation of $p(\lambda_i)$ from the characteristic function $\gamma(\mu, t)$ associated to the formulas $\text{dist}(x_1, x_2) \leq d$ by Lemma 2.

The clustering of our running example obtained from Theorem 7 is displayed on Fig. 8. Globular clusters are marked $M_1$, $M_2$, and $M_3$. The negligible sequence $S$ is depicted by big black square vertices. The residual sequence $R$ is what remains.

**Figure 8.** Clustering of our running example

**Remark 11.** Note that Theorem 7 cannot be extended to full first-order convergence. Consider the graph $G_n$ built by linking two complete graphs of order $n$ by a induced path of length $n$. This sequence is FO-convergent, and its two limit connected components with non-zero measure have measure $1/3$, and are formed by complete graph on a graph of measure $1/3$, to which is attached a ray. However, each of these connected components (taken alone) is not the first-order limit (and even not the elementary limit) of a sequence of finite graphs: the definable subgraph induced by vertices with degree two has a single vertex of degree one (in the induced subgraph), what cannot be the case in a finite graph.

A direct consequence of Theorem 7 stands in the following complete characterization of the globular clusters of a local convergent sequence.
**Theorem 8.** For a sequence $X$ of subsets of a local convergent sequence $A$ the following are equivalent

1. $X$ is a globular cluster of $A$;
2. there exists a negligible sequence $N$ and integers $i, j$ (with $1 \leq j \leq a_i$) such that for every integer $n$ it holds

$$X_n \Delta N_n \in \{G_n^{i,j,1}, G_n^{i,j,2}, \ldots, G_n^{i,j,b_i,n}\}.$$ 

**Proof.** If $X$ is obtained by interweaving clusters from $\{G_n^{i,j,1}, G_n^{i,j,2}, \ldots, G_n^{i,j,b_i,n}\}$, then $X$ is a cluster, which is obviously globular. Hence (2)$\Rightarrow$(1). Conversely, let $X$ be a globular cluster. As the partition is stable there exists, for every $\varepsilon > 0$, integers $i_0$ and $n_0$ such that for every $n \geq n_0$ it holds

$$\sum_{i > i_0} \sum_{j=1}^{a_i} \nu_A^n(Z_n^{i,j}) < \varepsilon.$$ 

Then notice that $X \cap R \approx 0$ as $R$ is residual and $X$ is not. According to Lemma 21, for each $i, j, k$ it either holds $X \cap G^{i,j,k} \approx 0$ or $X \not\approx G^{i,j,k}$. Let $n_1 \geq n_0$ be such that for every $n \geq n_0$ and every integers $i, j$ with $i \leq i_0$ and $X \cap Z^{i,j_n} \approx 0$ it holds

$$\nu_A^n(R_n \cap X_n) < \varepsilon$$ and $$\nu_A^n(R_n \cap Z_n^{i,j}) < \frac{\varepsilon}{\sum_{i=1}^{i_0} a_i}.$$ 

Then, letting $\varepsilon < \lim \nu_A^n(X)/4$ we get that there exists integers $i, j, k$ such that $X \not\approx G^{i,j,k}$. It follows that $X \not\approx G^{i',j',k'}$ if and only if $i = i'$ and $j = j'$. Thus for every $(i', j') \neq (i, j)$ it holds $X \cap Z^{i,j} \approx 0$, and thus it holds $\lim \inf \nu_A^n(X \cap Z^{i,j}) \geq 1 - 3\varepsilon$. Letting $\varepsilon \to 0$ we get that $Z^{i,j} \setminus X$ is negligible. As $X$ is globular and as $Z^{i,j}$ consists in connected components with same positive limit measure as $X$ selecting from $Z_n^{i,j}$ a connected component with maximal intersection with $X_n$ we get a globular cluster $Y$ such that $Y \approx X$. 

\[\square\]

10. **Conclusion and Future Work**

In this paper we have shown that a the local convergence of a sequence of finite structures is enough to obtain properties that cannot be expressed directly by means of a first-order formula: one can cluster the sequence into countably many globular clusters and a residual cluster. It is perhaps surprising that one can do so just from local convergence. The obtained clustering is natural and continuous. We believe that this analysis may be of interest in cluster analysis itself if only by the concepts that naturally arose in this study.

On the other hand, we feel that this is only the beginning of the story. Particularly because of their connection to expanders, we would like to further refine our
clustering and find further expanding (non globular) clusters. However this will require to consider a stronger notion of convergence, such as generalized local-global convergence. Our generalization of local-global convergence extends the notion of local-global convergence based on the colored neighborhood metric of Bollobás and Riordan [5], which was introduced by Hatami, Lovász, and Szegedy [11]. This will be the subject of a forthcoming paper.

ACKNOWLEDGMENTS

The authors would like to thank the referees for their careful reading of this manuscript and their valuable remarks and comments.

REFERENCES


JAROSLAV NEŠETŘIL, COMPUTER SCIENCE INSTITUTE OF CHARLES UNIVERSITY (IUUK AND ITI), MALOSTRANSKÉ NÁM.25, 11800 PRAHA 1, CZECH REPUBLIC

E-mail address: nesetril@kam.ms.mff.cuni.cz

PATRICE OSSONA DE MENDEZ, CENTRE D'ANALYSE ET DE MATHÉMATIQUES SOCIALES (CNRS, UMR 8557), 190-198 AVENUE DE FRANCE, 75013 PARIS, FRANCE AND COMPUTER SCIENCE INSTITUTE OF CHARLES UNIVERSITY (IUUK), MALOSTRANSKÉ NÁM.25, 11800 PRAHA 1, CZECH REPUBLIC

E-mail address: pom@ehess.fr