Cluster Analysis of Local Convergent Sequences of Structures

Jaroslav Nesetril, Patrice Ossona de Mendez

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

Date: October 26, 2015.

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 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 [8].

In this paper — which extends and precise some ideas introduced by the authors in [14] 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).

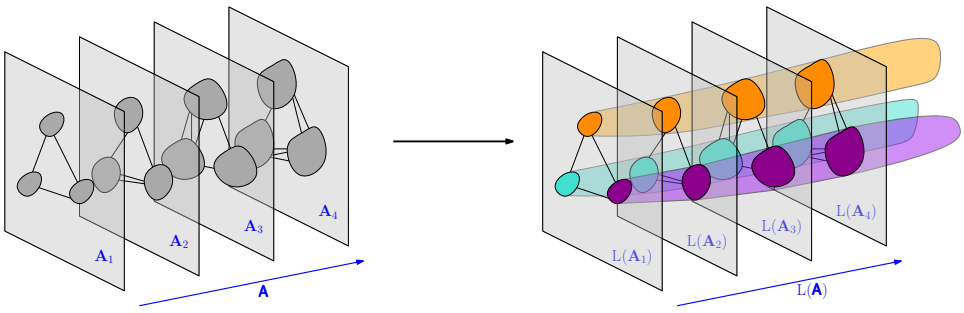

Figure 1. Segmentation of structures in a convergent sequence based on cluster analysis

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 [13], there is no restriction on the degrees of the considered graphs or structures.

Informally, a sequence $\mathbf{A} = (\mathbf{A}_n)_{n\in\mathbb{N}}$ of structures is *local convergent* if the probability $\langle \phi, \mathbf{A}_n \rangle$ (the *Stone pairing* of ϕ and $\mathbf{A}_n$) of satisfaction of every local first-order formula ϕ in structure $\mathbf{A}_n$ (for a random assignment of the free variables) converges as n grows to infinity. (Recall that a *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 ϕ. This can also be represented as a probability measure, as stated in the general representation theorem (Theorem 3), in a way extending Aldous-Hoover representation of left limits of dense graphs by infinite exchangeable graphs [1, 11] 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 the limit (although no *bona fide* limit structure is generally available). More: our cluster analysis will be a central tool to construct limit structures for sequences of graphs with locally few cycles (meaning that the number of cycles in the d-neighborhood of every vertex in every graph in the sequence is bounded by some fixed function of d). This will be the subject of a forthcoming paper [14].

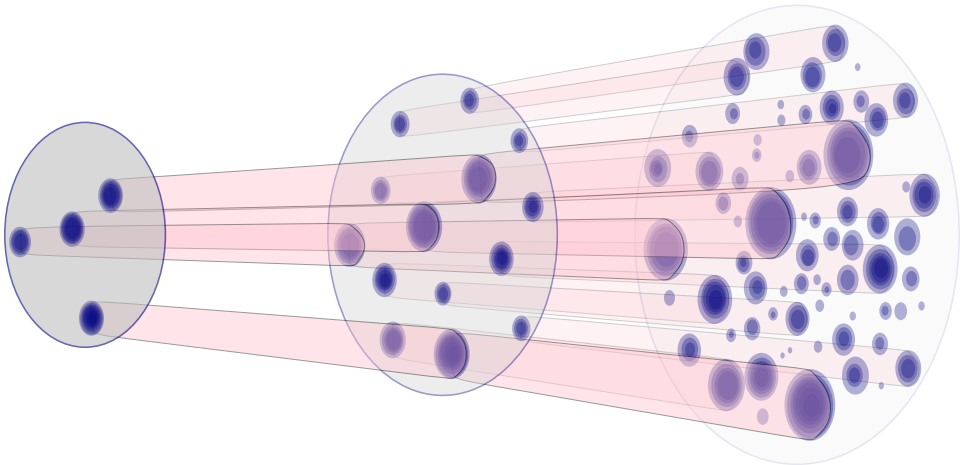

FIGURE 2. Typical shape of a structure continuously segmented by a clustering: dense spots correspond to globular clusters, and the background to the residual cluster. Biggest globular clusters appear first and then move apart from each other, while new (smaller) globular clusters appear and residual cluster becomes sparser and sparser.

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 $\mathbf{G} = (\mathbf{G}_n)_{i\in\mathbb{N}}$ 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

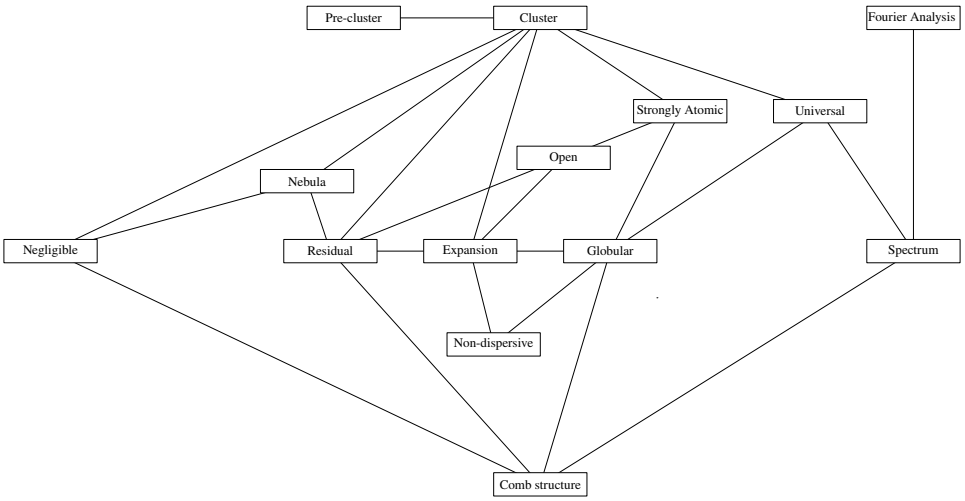

FIGURE 3. Semantic connections of new notions considered in this paper.

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 a 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 (see Fig 3): *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). 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* $\mathbf{A}$ *be a local convergent sequence of* σ*-structures. Then there exists a signature* σ^+ *obtained from* σ *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* $\mathbf{A}^+$ *of* $\mathbf{A}$ *with the following properties:*

- *For every* $i \in \mathbb{N}$*,* $\left(\bigcup_{j=1}^{N_i} M_{i,j}(\mathbf{A}_n^+)\right)_{n\in\mathbb{N}}$ *is a universal cluster;*
- *For every* $i \in \mathbb{N}$ *and every* $1 \leq j \leq N_i$*,* $\left(M_{i,j}(\mathbf{A}_n^+)\right)_{n\in\mathbb{N}}$ *is a globular cluster;*
- *Two clusters* $\left(M_{i,j}(\mathbf{A}_n^+)\right)_{n\in\mathbb{N}}$ *and* $\left(M_{i',j'}(\mathbf{A}_n^+)\right)_{n\in\mathbb{N}}$ *are interweaving if and only if* $i = i'$*;*
- $\left(M_R(\mathbf{A}_n^+)\right)_{n\in\mathbb{N}}$ *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 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 [6, 12] based on the notion of *left convergence*,
- the theory of bounded degree graph limits [3] based on the notion of *local convergence*.

These approaches have been (partly) unified by the authors in the setting of *structural limits* [13]. This last approach relies on a balance of model theoretic and functional analysis aspects. For a signature σ and a fragment X of the set of first-order formulas over the language generated by σ, we define for a finite σ-structure $\mathbf{A}$ and a formula $\phi \in X$ with free variables $x_1, \dots, x_p$ the *Stone pairing* of ϕ and $\mathbf{A}$ as

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

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

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

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

where $\nu_{\mathbf{A}}^{\otimes p}$ 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 σ will be denoted by $\mathrm{Rel}(\sigma)$.

In the following, we shall use the following convention (see Table 1):

- Structures are denoted by boldface capital letters $\mathbf{A}$;
- Sets are denoted by plain roman capital letters X, Y;
- Sequences of structures are denoted by boldface capital sans serif letter $\mathbf{A} = (\mathbf{A}_n)_{n\in\mathbb{N}}$;
- Sequence of sets are denoted by plain capital sans serif letter $\mathsf{X} = (X_n)_{n\in\mathbb{N}}$.

Let $\mathbf{A}$ be a σ-structure and let X be a subset of the domain A of $\mathbf{A}$. If $\nu_{\mathbf{A}}(X) > 0$ we define $\mathbf{A}[X]$ as the substructure of $\mathbf{A}$ induced by X endowed with probability measure defined by $\nu_{\mathbf{A}[X]}(Y) = \nu_{\mathbf{A}}(Y)/\nu_{\mathbf{A}}(X)$ (for every $Y \subseteq X$). If $\nu_{\mathbf{A}}(X) < 1$ we define $\mathbf{A} - X = \mathbf{A}[A \setminus X]$. We denote by $\mathrm{Gaif}(\mathbf{A})$ the Gaifman graph of $\mathbf{A}$. 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 $\mathbf{A}$ is $\mathrm{N}_{\mathbf{A}}(X)$. Consequently the set of all elements of A at distance at most d from an element of X is $\mathrm{N}_{\mathbf{A}}^d(X)$. The *outer vertex boundary* (or simply the *outer boundary*) of X in $\mathbf{A}$ is the set of

Symbol	Signification
$\mathrm{Rel}(\sigma)$	Set of all finite σ-structures (endowed with probability measure)
$\mathbf{A}$	A structure
A	The domain of structure $\mathbf{A}$
$\nu_{\mathbf{A}}$	Probability measure on the domain A of $\mathbf{A}$
X, Y	Subsets of A
$X \cap Y, X \cup Y, X \setminus Y$	Set operations
$X \subseteq Y$	Set inclusion
$\phi(\mathbf{A})$	Set of tuples satisfying ϕ in $\mathbf{A}$
$\langle \phi, \mathbf{A} \rangle$	Stone pairing of ϕ and $\mathbf{A}$
$\mathbf{A}[X]$	Substructure of $\mathbf{A}$ induced by X
$\mathbf{A} - X$	Substructure of $\mathbf{A}$ induced by $A \setminus X$
$\mathrm{N}^d_{\mathbf{A}}(X)$	Closed d-neighborhood of subset X in $\mathbf{A}$
$\partial_{\mathbf{A}}(X)$	Outer boundary of X in $\mathbf{A}$: $\partial_{\mathbf{A}}(X) = \mathrm{N}_{\mathbf{A}}(X) \setminus X$
$\mathbf{A}$	Sequence $(\mathbf{A}_n)_{n\in\mathbb{N}}$ of structures
A	Sequence $(A_n)_{n\in\mathbb{N}}$ of the domains of structures in $\mathbf{A}$
X	Sequence $(X_n)_{n\in\mathbb{N}}$ of subsets, cluster
$\mathsf{0}$	Sequence of empty sets: sequence X, where $X_n = \emptyset$
$\mathbf{A}_f$	Subsequence $(\mathbf{A}_{f(n)})_{n\in\mathbb{N}}$ of $\mathbf{A}$
X_f	Subsequence $(X_{f(n)})_{n\in\mathbb{N}}$ of X
$\nu_{\mathbf{A}}(\mathsf{X})$	Sequence $(\nu_{\mathbf{A}_n}(X_n))_{n\in\mathbb{N}}$ of measures of subsets
$\mathsf{X} \cap \mathsf{Y}, \mathsf{X} \cup \mathsf{Y}, \mathsf{X} \setminus \mathsf{Y}$	Sequences $(X_n \cap Y_n)_{n\in\mathbb{N}}$, $(X_n \cup Y_n)_{n\in\mathbb{N}}$, and $(X_n \setminus Y_n)_{n\in\mathbb{N}}$
$\mathsf{X} \subseteq \mathsf{Y}$	Pointwise sequence inclusion: $\mathsf{X} \subseteq \mathsf{Y} \iff (\forall n)\ X_n \subseteq Y_n$
$\mathbf{A}[\mathsf{X}]$	Sequence $(\mathbf{A}_n[X_n])_{n\in\mathbb{N}}$ of induced substructures
$\mathbf{A} - \mathsf{X}$	Sequence $(\mathbf{A}_n[A_n \setminus X_n])_{n\in\mathbb{N}}$ of induced substructures
$\mathrm{N}^d_{\mathbf{A}}(\mathsf{X})$	Sequence $(\mathrm{N}^d_{\mathbf{A}_n}(X_n))_{n\in\mathbb{N}}$ of closed d-neighborhoods
$\partial_{\mathbf{A}}\mathsf{X}$	Sequence $(\partial_{\mathbf{A}_n}(X_n))_{n\in\mathbb{N}}$ of outer boundaries
$\phi(\mathbf{A})$	Sequence $(\phi(\mathbf{A}_n))_{n\in\mathbb{N}}$ of satisfaction sets of ϕ
$\mathrm{L}_{\mathsf{X}}(\mathbf{A})$	Lifted sequence obtained by marking X in $\mathbf{A}$
$\lim \mathbf{A}$	Limit of $\mathbf{A}$ (as an abstract object)
$\langle \phi, \lim \mathbf{A} \rangle$	Limit Stone pairing: $\langle \phi, \lim \mathbf{A} \rangle = \lim_{n\to\infty} \langle \phi, \mathbf{A}_n \rangle$
Introduced in Section 8	
S_σ	Stone space associated to σ-structures
$P(S)$	Space of probability measures on space S
$\mathfrak{M}_\sigma$	Closure of the space of representation measures of finite σ-structures in $P(S_\sigma)$
$\mu_{\mathbf{A}}$	Representation measure of structure $\mathbf{A}$
$\mu_{\lim \mathbf{A}}$	Representation measure of the limit of sequence $\mathbf{A}$
$k(\phi)$	Function representing ϕ, s.t. $\langle \phi, \mathbf{A} \rangle = \int k(\phi)\, \mathrm{d}\mu_{\mathbf{A}}$
Introduced in Section 3	
$\phi \oplus \psi$	addition: $\phi \vee \psi$, defined if $\phi \wedge \psi = 0$
$\phi \ominus \psi$	subtraction: $\phi \wedge \neg\psi$, defined if $\phi \to \psi$
$\phi \otimes \psi$	free product of ϕ and ψ
Introduced in Section 4	
$\mathsf{X} \approx \mathsf{Y}$	Equivalent sequences ($\mathsf{X}\Delta\mathsf{Y}$ negligible)
Introduced in Section 5	
$\mathsf{X} \between \mathsf{Y}$	Interweaving clusters ($\lim \mathrm{L}_{\mathsf{X}}(\mathbf{A}) = \lim \mathrm{L}_{\mathsf{Y}}(\mathbf{A})$)

TABLE 1. Main symbols and notations of this paper

vertices in $A \setminus X$ with at least one neighbor in X [4]:

$$\partial_{\mathbf{A}} X = \mathrm{N}_{\mathbf{A}}(X) \setminus X.$$

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

Furthermore, we extend all operations defined on structures and subsets to sequences coordinate-wise: The sequence $\mathbf{A}$ has domain A (meaning $\mathbf{A}_n$ has domain A_n); for $\mathsf{X} \subseteq \mathsf{A}$ (meaning $X_n \subseteq A_n$) we denote by $\mathbf{A}[\mathsf{X}]$ the sequence $(\mathbf{A}_n[X_n])_{n\in\mathbb{N}}$, by $\partial_{\mathbf{A}}\mathsf{X}$ the sequence $(\partial_{\mathbf{A}_n} X_n)_{n\in\mathbb{N}}$, by $\mathsf{X} \subseteq \mathsf{Y}$ the inclusions $X_n \subseteq Y_n$, by $\phi(\mathbf{A})$ the sequence of the sets $\phi(\mathbf{A}_n)$, by $\nu_{\mathbf{A}}(\mathsf{X})$ the sequence of the measures $\nu_{\mathbf{A}_n}(X_n)$, etc. Also, for increasing $f : \mathbb{N} \to \mathbb{N}$ we denote by $\mathbf{A}_f$ the subsequence $(\mathbf{A}_{f(n)})_{n\in\mathbb{N}}$ of $\mathbf{A}$ and by X_f the subsequence $(X_{f(n)})_{n\in\mathbb{N}}$ of X. For instance, $\mathbf{A}$ denotes a sequence of structures whose nth term is $\mathbf{A}_n$, and A denotes the sequence of the domains A_n of the structures $\mathbf{A}_n$.

Definition 1. For σ-structures $\mathbf{A}_1, \mathbf{A}_2, \dots$ and not negative reals $\lambda_1, \lambda_2, \dots$ with sum 1 we define $\sum_i \lambda_i \mathbf{A}_i$ as the σ-structure $\mathbf{A}$ obtained by endowing the disjoint union of the σ-structures $\mathbf{A}_1, \mathbf{A}_2, \dots$ with the probability measure $\nu_{\mathbf{A}} = \sum_i \lambda_i \nu_{\mathbf{A}_i}$.

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

To deal marking we introduce the following notion of lift:

Definition 2. Let $\sigma \subset \sigma^+$ be countable signatures, let $\mathbf{A}$ be a sequence of σ-structures, and let $\mathbf{B}$ be a sequence of σ^+-structures.

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

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

For the benefit of the reader we included in Table 1 a 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 ϕ is *local* if there is some integer r such that the satisfaction of ϕ only depends on the distance r neighborhood of its free variables. Let $\mathrm{FO}^{\mathrm{local}}(\sigma)$ be the fragment of local first-order formulas (for given signature σ). The following is the key definition.

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

Note that for bounded degree graphs our notion of local convergence is equivalent to the notion of local convergence introduced in [3] (see [13]). For general graphs (or regular hypergraphs), local convergence is stronger than the left convergence considered by [12, 7].

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

Fact 1. Let X, Y be subsets of the domain A of a structure $\mathbf{A}$, 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

$$\mathrm{N}^d_{\mathbf{A}}(\partial_{\mathbf{A}} Z) \subseteq \mathrm{N}^{d+1}_{\mathbf{A}}(\partial_{\mathbf{A}} X) \cup \mathrm{N}^{d+1}_{\mathbf{A}}(\partial_{\mathbf{A}} Y).$$

A first-order formula ϕ with free variables $x_1, \ldots, x_p$ is *r-local* if, for every structure $\mathbf{A}$ and elements $v_1, \ldots, v_p \in A$ it holds

$$\mathbf{A} \models \phi(v_1, \ldots, v_p) \quad \iff \quad \mathbf{A}[\mathrm{N}^r_{\mathbf{A}}(\{v_1, \ldots, v_p\})] \models \phi(v_1, \ldots, v_p).$$

A formula is called *local* if it is r-local for some r. The set of all local first-order formulas (in the language of the considered signature) is denoted by $\mathrm{FO}^{\mathrm{local}}$, and we simply use the term of *local convergence* for $\mathrm{FO}^{\mathrm{local}}$-convergence. Note that this notion of convergence extends Benjamini-Schramm's notion of local convergence:

Fact 2 ([13]). 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 $\mathrm{FO}^{\mathrm{local}}$-convergence).

Indeed, by Gaifman locality theorem [9], for every local formula ϕ 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.

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 ϕ with free variables $x_1, \ldots, x_p$ is *strongly r-local* if it is r-local and the following entailment holds:

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

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 ϕ is *packed* if its free variables are $x_1, \ldots, x_p$ (for some $p \in \mathbb{N}$). For a formula ϕ with free variables $x_{i_1}, \ldots, x_{i_p}$ and an injection $\iota : \mathbb{N} \to \mathbb{N}$, $\iota(\phi)$ denotes the formula ϕ where all the occurrences of x_j are replaced by $x_{\iota(j)}$. We denote by τ be the injection $i \mapsto i + 1$.

Definition 4. A *weak algebra* of formula is a set $\mathcal{S}$ of (logical equivalence class of) formulas which is closed under the following (partially defined) operations:

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

Note that for every ϕ, ψ, we have:

- If $\phi \oplus \psi$ is defined then for every structure $\mathbf{A}$ it holds

$$(\phi \oplus \psi)(\mathbf{A}) \equiv \phi(\mathbf{A}) \cup \psi(\mathbf{A})$$
$$\langle \phi \oplus \psi, \mathbf{A} \rangle = \langle \phi, \mathbf{A} \rangle + \langle \psi, \mathbf{A} \rangle$$

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

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 ι of $[p+q]$ such that
$$\iota(X \times A^q) = Y \times A^p.$$
Also, note that if $\phi \oplus \psi$ is defined then $\phi = (\phi \oplus \psi) \ominus \psi$.

Theorem 2. *The smallest weak algebra containing all strongly local formulas is the weak algebra of all local formulas.*

Proof. One direction is obvious (as local formulas form a weak algebra). For the other direction, consider an r-local formula ϕ with free variables x_i for $i \in I$. Let $\mathfrak{F}_I$ be the set of all graphs with vertex set I. Obviously it holds
$$\phi = \bigoplus_{F \in \mathfrak{F}_I} \left(\bigwedge_{ij \in E(F)} (\mathrm{dist}(x_i, x_j) \leq 2r) \wedge \bigwedge_{ij \notin E(F)} (\mathrm{dist}(x_i, x_j) > 2r) \wedge \phi \right).$$
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 ϕ has the following form:
$$\phi = \bigwedge_{ij \in E(F)} (\mathrm{dist}(x_i, x_j) \leq 2r) \wedge \bigwedge_{ij \notin E(F)} (\mathrm{dist}(x_i, x_j) > 2r) \wedge \bigwedge_{z=1}^{k} \rho_z,$$
where ρ_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
$$\phi_z = \bigwedge_{ij \in E(F_z)} (\mathrm{dist}(x_i, x_j) \leq 2r) \wedge \rho_z.$$
Then it holds
$$\bigotimes_{z=1}^{k} \phi_z = \bigoplus_{H \in \mathfrak{F}'} \left(\bigwedge_{ij \in E(H)} (\mathrm{dist}(x_i, x_j) \leq 2r) \wedge \bigwedge_{ij \notin E(H)} (\mathrm{dist}(x_i, x_j) > 2r) \wedge \bigwedge_{z=1}^{k} \rho_z \right),$$
where $\mathfrak{F}'$ 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 $\mathfrak{F}'$ 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
$$\psi = \bigoplus_{H \in \mathfrak{F}' \setminus \{F\}} \left(\bigwedge_{ij \in E(H)} (\mathrm{dist}(x_i, x_j) \leq 2r) \wedge \bigwedge_{ij \notin E(H)} (\mathrm{dist}(x_i, x_j) > 2r) \wedge \bigwedge_{z=1}^{k} \rho_z \right),$$

it holds

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

By induction, ψ belongs to the weak algebra generated by strongly local formulas, hence so does ϕ. □

We now take time for three important corollaries.

Corollary 1. *For every r-local formula ϕ with p free variables, there exist finitely many $(2pr)$-strongly local formulas ϕ_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 $\mathbf{A}$ it holds*

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

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 $\mathbf{A}$ is an infinite structure, whose domain is a measurable space. If, for every strongly local formula ϕ the set $\phi(\mathbf{A})$ is measurable (with respect to product measure) then for every first-order formula ψ the set $\psi(\mathbf{A})$ is measurable (with respect to product measure).*

Proof. This is a direct consequence of Theorem 2 and Gaifman locality theorem [9]. □

Corollary 3. *A sequence $\mathbf{A}$ 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 justifies to relate it to the notion of lift introduced in Section 2.

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 $\mathbf{A}$ be a structure, let $d \in \mathbb{N}$ and let $\epsilon > 0$. A subset $X \subset A$ of elements of $\mathbf{A}$ is (d, ϵ)-*negligible* in $\mathbf{A}$ if

$$\nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(X)) < \epsilon.$$

The main property of (d, ϵ)-negligible sets is the following:

Lemma 1. *Let $\phi \in \mathrm{FO}^{\mathrm{local}}_p$ be r-local with $r < d$, and let X be a (d, ϵ)-negligible set of a structure $\mathbf{A}$. Then*

$$|\langle \phi, \mathbf{A} \rangle - \langle \phi, \mathbf{A} - X \rangle| < 2p\epsilon.$$

Moreover, suppose $\mathbf{B}$ is a structure with same domain as $\mathbf{A}$ such that $\mathbf{A} - X = \mathbf{B} - X$ then

$$|\langle \phi, \mathbf{A} \rangle - \langle \phi, \mathbf{B} \rangle| < p\epsilon.$$

Proof. We first prove the second inequality.

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

$$0 \le \langle \phi, \mathrm{L}(\mathbf{A})\rangle - \langle \phi \wedge \neg\psi, \mathrm{L}(\mathbf{A})\rangle < \langle \psi, \mathrm{L}(\mathbf{A})\rangle < 1-(1-\epsilon)^p < p\epsilon$$
$$0 \le \langle \phi, \mathrm{L}(\mathbf{B})\rangle - \langle \phi \wedge \neg\psi, \mathrm{L}(\mathbf{B})\rangle < \langle \psi, \mathrm{L}(\mathbf{B})\rangle < p\epsilon$$

Thus, as $\phi(\mathrm{L}(\mathbf{A})) = \phi(\mathbf{A})$, $\phi(\mathrm{L}(\mathbf{B})) = \phi(\mathbf{B})$ and $(\phi\wedge\neg\psi)(\mathrm{L}(\mathbf{A})) = (\phi\wedge\neg\psi)(\mathrm{L}(\mathbf{B}))$ it holds

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

For the second inequality, we have likewise

$$0 \le \langle \phi, \mathrm{L}(\mathbf{A}) - X\rangle - \langle \phi \wedge \neg\psi, \mathrm{L}(\mathbf{A}) - X\rangle < \langle \psi, \mathrm{L}(\mathbf{A}) - X\rangle < \frac{p\epsilon}{\nu_{\mathbf{A}}(A\setminus X)^p}$$

Moreover, as ϕ is r-local, it holds $(\phi \wedge \neg\psi)(\mathrm{L}(\mathbf{A})) = (\phi \wedge \neg\psi)(\mathrm{L}(\mathbf{A}) - X)$ hence

$$\langle \phi \wedge \neg\psi, \mathrm{L}(\mathbf{A})\rangle = \nu_{\mathbf{A}}(A\setminus X)^p \langle \phi \wedge \neg\psi, \mathrm{L}(\mathbf{A}) - X\rangle.$$

Thus

$$|\langle \phi, \mathrm{L}(\mathbf{A})\rangle - \nu_{\mathbf{A}}(A\setminus X)^p \langle \phi, \mathrm{L}(\mathbf{A}) - X\rangle| < p\epsilon$$

Hence

$$\begin{aligned}|\langle \phi, \mathrm{L}(\mathbf{A})\rangle - \langle \phi, \mathrm{L}(\mathbf{A}) - X\rangle| &\le |\langle \phi, \mathrm{L}(\mathbf{A})\rangle - \nu_{\mathbf{A}}(A\setminus X)^p \langle \phi, \mathrm{L}(\mathbf{A}) - X\rangle| + 1 - \nu_{\mathbf{A}}(A\setminus X)^p \\ &< p\epsilon + 1 - (1-\epsilon)^p \\ &< 2p\epsilon.\end{aligned}$$

hence the result, as $\langle \phi, \mathrm{L}(\mathbf{A})\rangle = \langle \phi, \mathbf{A}\rangle$ and $\langle \phi, \mathrm{L}(\mathbf{A}) - X\rangle = \langle \phi, \mathbf{A} - X\rangle$. □

Definition 6. A (d,ϵ)-*fragmentation* of a structure $\mathbf{A}$ is a (at most) countable partition $(S, X_1, X_2, \dots)$ of A such that no element in X_i has a neighbor in X_j for $i \neq j$ and S is (d,ϵ)-negligible in $\mathbf{A}$.

Lemma 2. *Assume $(S, X_1, X_2, \dots)$ is a (d,ϵ)-fragmentation of $\mathbf{A}$ and let ϕ be a strongly r-local formula ($r \le d$) with free variables $x_1,\dots,x_p$. Then*

$$\left|\langle \phi, \mathbf{A}\rangle - \sum_{i\ge 1} \nu_{\mathbf{A}}(X_i)^p \langle \phi, \mathbf{A}[X_i]\rangle\right| < 2p\epsilon.$$

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

We now consider how the notion of (d,ϵ)-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 $\mathsf{X} \subseteq \mathsf{A}$ is *negligible* in A if

$$\forall d \in \mathbb{N}: \quad \limsup_{n\to\infty} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(X_n)) = 0,$$

what we simply formulate as

$$\forall d \in \mathbb{N}: \quad \limsup \nu_{\mathsf{A}}(\mathrm{N}^d_{\mathsf{A}}(\mathsf{X})) = 0.$$

Two sequences X and Y of subsets are *equivalent* in A (and we note $\mathsf{X} \approx \mathsf{Y}$ if the sequence $\mathsf{X}\,\Delta\,\mathsf{Y} = (X_n\,\Delta\,Y_n)_{n\in\mathbb{N}}$ is negligible in A.

We denote by $\mathsf{0}$ the sequence of empty subsets. Hence $\mathsf{X} \approx \mathsf{0}$ is equivalent to the property that X is negligible.

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

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

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

Lemma 3. *Let $\mathbf{A}$ and $\mathbf{B}$ be equivalent sequences of structures.*

Then $\mathbf{A}$ is local-convergent if and only if $\mathbf{B}$ 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".

5.1. **Clusters.** In our setting, where clustering is performed on a local convergent sequence $\mathbf{A}$, 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 $\mathbf{A}$ be a local-convergent sequence of structures.

A sequence $\mathsf{X} \subseteq \mathsf{A}$ is a *cluster* of $\mathbf{A}$ if the following conditions hold:

(1) the lifted sequence $\mathrm{L}_{\mathsf{X}}(\mathbf{A})$ obtained by marking set X_n in $\mathbf{A}_n$ by a new unary relation M_{X} is local-convergent;
(2) the sequence $\partial_{\mathbf{A}}\mathsf{X}$ is negligible in $\mathbf{A}$.

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 $\mathbf{A}$ such that $\mathsf{Y} \preceq \mathsf{X}$ either $\mathsf{Y} \approx \mathsf{0}$ or $\mathsf{Y} \approx \mathsf{X}$; the cluster X is *strongly atomic* if X_f is an atomic cluster of $\mathbf{A}_f$ 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 $\mathbf{A}_f$ with $\mathsf{Y}_f \subseteq \mathsf{X}_f$ is trivial. Finally, a cluster X is *universal* for $\mathbf{A}$ if X is a cluster of every conservative lift of $\mathbf{A}$.

Lemma 4. *Let X be a cluster of $\mathbf{A}$ and let Y be a sequence of subsets. Then $\mathsf{X} \approx \mathsf{Y}$ in $\mathbf{A}$ if and only if Y is a cluster of $\mathbf{A}$ and*

$$\limsup \nu_{\mathbf{A}}(\mathsf{X} \,\Delta\, \mathsf{Y}) = 0.$$

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

$$\begin{aligned}\mathrm{N}^d_{\mathbf{A}_n}(X_n \,\Delta\, Y_n) &\subseteq (X_n \,\Delta\, Y_n) \cup \mathrm{N}^d_{\mathbf{A}_n}(\partial_{\mathbf{A}_n}(X_n \,\Delta Y_n)) \\ &\subseteq (X_n \,\Delta\, Y_n) \cup \mathrm{N}^d_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} X_n) \cup \mathrm{N}^d_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} Y_n)\end{aligned}$$

Thus $\mathsf{X} \,\Delta\, \mathsf{Y}$ is negligible in $\mathbf{A}$, that is $\mathsf{X} \approx \mathsf{Y}$.

Conversely, assume $\mathsf{X} \approx \mathsf{Y}$. Then obviously $\limsup \nu_{\mathbf{A}}(\mathsf{X} \,\Delta\, \mathsf{Y}) = 0$. As $\partial_{\mathbf{A}_n} Y_n \subseteq \partial_{\mathbf{A}_n} X_n \cup \mathrm{N}_{\mathbf{A}_n}(X_n \,\Delta\, Y_n)$, and as $X \Delta Y$ is negligible since $X \approx Y$, the sequence $\partial_{\mathbf{A}}\mathsf{Y}$

is negligible. Moreover, as $\mathrm{L}_{\mathsf{X}}(\mathbf{A}) \approx \mathrm{L}_{\mathsf{Y}}(\mathbf{A})$ (considering we use the same symbol for both lifts), we deduce that Y is a cluster of $\mathbf{A}$. □

In particular, if X is a cluster and $\mathsf{Y} \approx \mathsf{X}$ then Y is a cluster.

We have the following alternative characterization of clusters:

Lemma 5. *Let $\mathbf{A}$ be a local-convergent sequence of structures.*

A sequence $\mathsf{X} \subseteq \mathsf{A}$ is a cluster of $\mathbf{A}$ if either $\mathsf{X} \approx \mathsf{0}$ or the following conditions hold:

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

Proof. Assume X is a cluster of $\mathbf{A}$, and let $\alpha = \langle M_{\mathsf{X}}, \lim \mathrm{L}_{\mathsf{X}}(\mathbf{A})\rangle$. If $\alpha = 0$ then X is negligible in $\mathbf{A}$ as for every $d, n \in \mathbb{N}$ it holds $\mathrm{N}^d_{\mathbf{A}_n}(X_n) = X_n \cup \mathrm{N}^{d-1}_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} X_n)$ and thus

$$\limsup \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\mathsf{X})) \leq \lim \nu_{\mathbf{A}}(\mathsf{X}) + \limsup \nu_{\mathbf{A}}(\mathrm{N}^{d-1}_{\mathbf{A}}(\partial_{\mathbf{A}}\mathsf{X}) = \alpha = 0.$$

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

$$\begin{aligned}\langle \phi, \mathbf{A}_n[X_n]\rangle &= \langle \phi | M_{\mathbf{X}}, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)[X_n]\rangle \\ &= \alpha^{-p}\langle \phi | M_{\mathbf{X}}, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n) - \partial_{\mathbf{A}_n}(X_n)\rangle + o(1) \\ &= \alpha^{-p}\langle \phi | M_{\mathbf{X}}, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)\rangle + o(1)\end{aligned}$$

It follows that the sequence $\mathbf{A}[\mathsf{X}]$ is local-convergent.

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

$$\begin{aligned}\langle \phi, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)\rangle &= \langle \phi, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n) - \partial_{\mathbf{A}_n} X_n\rangle + o(1) \\ &= \alpha^p \langle \phi, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)[X_n]\rangle + (1-\alpha)^p \langle \phi, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)[A_n \setminus X_n \setminus \partial_{\mathbf{A}_n} X_n]\rangle + o(1) \\ &= \alpha^p \langle \phi^+, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)[X_n]\rangle + (1-\alpha)^p \langle \phi^-, \mathrm{L}_{\mathsf{X}}(\mathbf{A}_n)[A_n \setminus X_n \setminus \partial_{\mathbf{A}_n} X_n]\rangle + o(1) \\ &= \alpha^p \langle \phi^+, \mathbf{A}_n[X_n]\rangle + (1-\alpha)^p \langle \phi^-, \mathbf{A}_n[A_n \setminus X_n \setminus \partial_{\mathbf{A}_n} X_n]\rangle + o(1) \\ &= \alpha^p \langle \phi^+, \mathbf{A}_n[X_n]\rangle + \langle \phi^-, \mathbf{A}_n\rangle - \alpha^p \langle \phi^-, \mathbf{A}_n[X_n]\rangle + o(1)\end{aligned}$$

Hence $\mathrm{L}_{\mathsf{X}}(\mathbf{A})$ is local convergent. □

Definition 9. Two clusters X and Y of a local-convergent sequence $\mathbf{A}$ are *interweaving*, and we note $\mathsf{X} \between \mathsf{Y}$ if every sequence Z with $Z_n \in \{X_n, Y_n\}$ is a cluster of $\mathbf{A}$.

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 $\mathbf{A}$. The following are equivalent:*

1. *X and Y are interweaving;*
2. *$\lim \mathrm{L}_{\mathsf{X}}(\mathbf{A}) = \lim \mathrm{L}_{\mathsf{Y}}(\mathbf{A})$;*
3. *either $\mathsf{X} \approx \mathsf{Y} \approx \mathsf{0}$ or the following two conditions hold:*
   (a) *$\lim \mathbf{A}[\mathsf{X}] = \lim \mathbf{A}[\mathsf{Y}]$;*
   (b) *$\lim \nu_{\mathbf{A}}(\mathsf{X}) = \lim \nu_{\mathbf{A}}(\mathsf{Y})$.*

Proof. Let us prove (1)⇒(2)⇒(3)⇒(1).

(1)⇒(2): Let Z_n be X_n if n is odd and Y_n if n is even. As $\mathsf{X} \between \mathsf{Y}$, the sequence Z is a cluster and (by considering the common subsequences) it holds $\lim \mathrm{L}_{\mathsf{X}}(\mathbf{A}) = \lim \mathrm{L}_{\mathsf{Z}}(\mathbf{A}) = \lim \mathrm{L}_{\mathsf{Y}}(\mathbf{A})$.

(2)⇒(3): Let $\alpha = \lim\langle M(x_1), \mathrm{L}_{X_n}(\mathbf{A}_n)\rangle$. Then either $\alpha = 0$ and $\mathsf{X} \approx \mathsf{Y} \approx \mathsf{0}$ or $\lim \nu_{\mathbf{A}}(\mathsf{X}) = \lim \nu_{\mathbf{A}}(\mathsf{Y})\alpha > 0$. In the later case, for any local formula ϕ, let $\widetilde{\phi}$ be the formula where all the variables (free or bound) are constrained to belong to relation M. For sufficiently large n (so that $\langle M(x_1), \mathrm{L}_{X_n}(\mathbf{A}_n)\rangle > 0$ and $\langle M(x_1), \mathrm{L}_{Y_n}(\mathbf{A}_n)\rangle > 0$) it holds

$$\langle \phi, \mathbf{A}_n[X_n]\rangle = \frac{\langle \widetilde{\phi}, \mathrm{L}_{X_n}(\mathbf{A}_n)\rangle}{\langle M(x_1), \mathrm{L}_{X_n}(\mathbf{A}_n)\rangle} \text{ and } \langle \phi, \mathbf{A}_n[Y_n]\rangle = \frac{\langle \widetilde{\phi}, \mathrm{L}_{Y_n}(\mathbf{A}_n)\rangle}{\langle M(x_1), \mathrm{L}_{Y_n}(\mathbf{A}_n)\rangle}.$$

Thus $\lim \mathbf{A}[\mathsf{X}] = \lim \mathbf{A}[\mathsf{Y}]$.

(3)⇒(1): If $\mathsf{X} \approx \mathsf{Y} \approx \mathsf{0}$ then obviously $\mathsf{X} \between \mathsf{Y}$. Otherwise, consider arbitrary Z with $Z_n \in \{X_n, Y_n\}$. Assume for contradiction that $\mathbf{A}[\mathsf{Z}$ does not converge. Then we can extract two subsequences Z_f and Z_g such that $\lim \mathbf{A}_f[\mathsf{Z}_f \neq \lim \mathbf{A}_g[\mathsf{Z}_g]$. 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 $\lim \mathbf{A}[\mathsf{X}] = \lim \mathbf{A}[\mathsf{Y}]$ we get a contradiction, so $\mathbf{A}[\mathsf{Z}$ converges. Similarly, $\nu_{\mathbf{A}}(\mathsf{Z})$ converges. As $\partial_{\mathbf{A}_n}(Z_n) \subseteq \partial_{\mathbf{A}_n}(X_n) \cup \partial_{\mathbf{A}_n}(Y_n)$ and $\partial_{\mathbf{A}}\mathsf{X} \approx \partial_{\mathbf{A}}\mathsf{Y} \approx \mathsf{0}$ we get $\partial_{\mathbf{A}}\mathsf{Z} \approx \mathsf{0}$. Altogether, this means that Z is a cluster of $\mathbf{A}$.

□

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 $\mathbf{A}$ be a local-convergent sequence, and let X be a cluster of $\mathbf{A}$. Then $\mathsf{Y} = \mathsf{A} \setminus \mathsf{X}$ is a cluster of $\mathbf{A}$.*

Proof. First notice that for every integer d it holds

$$\mathrm{N}^d_{\mathbf{A}}(\partial_{\mathbf{A}}(\mathsf{Y})) \subseteq \mathrm{N}^{d+1}_{\mathbf{A}}(\partial_{\mathbf{A}}(\mathsf{X}))$$

thus $\lim \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\partial_{\mathbf{A}}\mathsf{Y})) = 0$, that is $\mathsf{Y} \approx \mathsf{0}$. As $L_{\mathsf{Y}}(\mathbf{A})$ can be obtained from $L_{\mathsf{X}}(\mathbf{A})$ by taking for M_{Y} the negation of M_{X} it is clear that $L_{\mathsf{Y}}(\mathbf{A})$ is local convergent. □

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

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

$$\mathbf{A}_n = \begin{cases} \mathbf{E}_{5n} \cup \mathbf{E}_{6n} \cup \mathbf{E}_{8n} & \text{if } n \text{ is odd} \\ \mathbf{E}_{2n} \cup \mathbf{E}_{3n} \cup \mathbf{E}_{4n} \cup \mathbf{E}_{10n} & \text{if } n \text{ is even} \end{cases}$$

Then it is easily checked that $\mathbf{A}$ is local convergent, and that the only clusters of $\mathbf{A}$ are (up to equivalence) $\mathsf{0}, \mathsf{X}, \mathsf{Y}, \mathsf{A} \setminus \mathsf{X}, \mathsf{A} \setminus \mathsf{Y}$, 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 $\mathbf{A}_n$ 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 A be a local convergent sequence, and let $\mathsf{C}^1, \ldots, \mathsf{C}^i, \ldots$ be countably many clusters of A. Then the following conditions are equivalent:*
(1) *The lifted sequence $\mathrm{L}(\mathsf{A})$ 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)⇒(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)⇒(1). According to Corollary 3, in order to prove (1) it is sufficient to prove that for every strongly local formula ϕ (with some p free variables) the sequence $\langle \phi, \mathrm{L}(\mathsf{A})\rangle$ converges. Let N be the maximum index of a mark symbol used in ϕ. For $I \subseteq [N]$ denote by ϕ^I the formula where every term $M_i(x)$ is replaced by `true` if $i \in I$ and `false` if $i \notin I$. Define θ_I as the formula $\bigwedge_{i\in I} M_i(x_1) \wedge \bigwedge_{i\notin I} \neg M_i(x_1)$. Let $S = \bigcup_{I\subseteq[N]} \partial_{\mathsf{A}}\theta_I(\mathsf{A})$. As each $\theta_I(\mathsf{A})$ defines a cluster, S is negligible. Thus Then it holds

$$\begin{aligned}\langle \phi, \mathrm{L}(\mathbf{A}_n)\rangle &= \langle \phi, \mathrm{L}(\mathbf{A}_n) - S\rangle + o(1)\\ &= \sum_{I\subseteq[N]} \nu_{\mathbf{A}_n}(\theta_I(\mathbf{A}_n))^p \, \langle \phi^I, \mathbf{A}_n[\theta_I(\mathbf{A}_n)]\rangle + o(1)\end{aligned}$$

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

That (2)⇒(3) is trivial. Let us now prove (3)⇒(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 $\mathsf{X}\cap\mathsf{Y}$ are all clusters of A then so is $\mathsf{X}\setminus\mathsf{Y}$. To prove this, let $\mathsf{S} = \partial_{\mathsf{A}}\mathsf{X} \cup \partial_{\mathsf{A}}(\mathsf{X}\cap\mathsf{Y})$. Then S is negligible and thus for every strongly local formula ϕ (with p free variables) it holds

$$\begin{aligned}\langle \phi, \mathrm{L}_{X\setminus Y}(\mathbf{A}_n)\rangle = \langle \phi^0, \mathbf{A}_,\rangle &+ \nu_{\mathbf{A}_n}(X_n)^p \left(\langle \phi^1, \mathbf{A}_n[X_n]\rangle - \langle \phi^0, \mathbf{A}_n[X_n]\rangle\right)\\ &+ \nu_{\mathbf{A}_n}(X_n\cap Y_n)^p \left(\langle \phi^0, \mathbf{A}_n[X_n\cap Y_n]\rangle - \langle \phi^1, \mathbf{A}_n[X_n\cap Y_n]\rangle\right) + o(1),\end{aligned}$$

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

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

Then the lifted sequence $\mathrm{L}(\mathsf{A})$ 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 σ-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, namely:

(1) The sequence $\bigcup_i \partial_{\mathbf{A}} \mathsf{C}^i$ is negligible;
(2) The clusters C^i form a stable partition of A in the sense that $\sum_i \lim \nu_{\mathbf{A}}(\mathsf{C}^i) = 1$.

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

Proof. Let $S = \bigcup_i \partial_{\mathbf{A}} \mathsf{C}^i$. Let ϕ 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 $\nu_{\mathbf{A}}(\mathrm{N}_{\mathbf{A}}^{r+1}(S)) < \epsilon/8p$. Let $I \subseteq \mathbb{N}$ and let $\mathrm{L}_I(\mathbf{A})$ be the sequence obtained from $\mathbf{A}$ by marking elements of $\bigcup_{i\in I} \mathsf{C}^i$ by a new mark M. Let ψ be an r-local strongly local formula in the extended signature, and let ψ^0 (resp. ψ^1) be the formula obtained from ψ 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

$$\left|\langle \psi, \mathrm{L}_I(\mathbf{A}_n)\rangle - \sum_{i\notin I} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \psi^0, \mathbf{A}_n[C_n^i]\rangle - \sum_{i\in I} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \psi^1, \mathbf{A}_n[C_n^i]\rangle\right| < \epsilon/4.$$

As $\sum_{i\geq 1} \lim \nu_{\mathbf{A}}(\mathsf{C}^i) = 1$ there exists $i_0 \in \mathbb{N}$ such that $\sum_{i>i_0} \lim \nu_{\mathbf{A}}(\mathsf{C}^i) < \epsilon/8$ thus

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

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

$$\left|\nu_{\mathbf{A}_n}(C_n^i)^p \langle \psi^k, \mathbf{A}_n[C_n^i]\rangle - \left(\lim \nu_{\mathbf{A}}(\mathsf{C}^i)\right)^p \lim\langle \psi^k, \mathbf{A}[\mathsf{C}^i]\rangle\right| < \epsilon/4i_0$$

and

$$\left|\nu_{\mathbf{A}_n}(C_n^i) - \lim \nu_{\mathbf{A}}(\mathsf{C}^i)\right| < \epsilon/4i_0.$$

We deduce that

$$\sum_{i>i_0} \nu_{\mathbf{A}_n}(C_n^i) = 1 - \sum_{i=1}^{i_0} \nu_{\mathbf{A}_n}(C_n^i) < 1 - \sum_{i=1}^{i_0} \lim \nu_{\mathbf{A}}(\mathsf{C}^i) + \epsilon/4 = \sum_{i>i_0} \lim \nu_{\mathbf{A}}(\mathsf{C}^i) + \epsilon/4 < 3\epsilon/8.$$

From this follows that

$$\left|\langle \psi, \mathrm{L}_I(\mathbf{A}_n)\rangle - \sum_{i\notin I} \left(\lim \nu_{\mathbf{A}}(\mathsf{C}^i)\right)^p \lim\langle \psi^0, \mathbf{A}[\mathsf{C}^i]\rangle - \sum_{i\in I} \left(\lim \nu_{\mathbf{A}}(\mathsf{C}^i)\right)^p \lim\langle \psi^1, \mathbf{A}[\mathsf{C}^i]\rangle\right| < \epsilon.$$

It follows that $\mathrm{L}_I(\mathbf{A})$ is local convergent. As $\partial_{\mathbf{A}}\left(\bigcup_{i\in I} \mathsf{C}^i\right) \subseteq \bigcup_i \partial_{\mathbf{A}} \mathsf{C}^i$ is negligible by assumption, we deduce that $\bigcup_{i\in I} \mathsf{C}^i$ is a cluster. □

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* $\mathbf{A}$*, and let* $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$ *be the lift of* $\mathbf{A}$ *obtained by marking* C *by a new unary symbol* M_{C}*.*

Then, a sequence X *is a cluster of* $\mathbf{A}$ *if and only if it is a cluster of* $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$*.*

Proof. Of course, every cluster of $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$ is a cluster of $\mathbf{A}$.

Assume X is a cluster of $\mathbf{A}$. Then, by definition, the sequence $\mathrm{L}_{\mathsf{X}}(\mathbf{A})$ is a local convergent lift of $\mathbf{A}$. As $\mathbf{C}$ is universal, it is a cluster of $\mathrm{L}_{\mathsf{X}}(\mathbf{A})$ hence the sequence $\mathrm{L}_{\mathsf{C}}(\mathrm{L}_{\mathsf{X}}(\mathbf{A}))$ is local convergent. As $\mathrm{L}_{\mathsf{C}}(\mathrm{L}_{\mathsf{X}}(\mathbf{A})) = \mathrm{L}_{\mathsf{X}}(\mathrm{L}_{\mathsf{C}}(\mathbf{A}))$ we deduce that X is a cluster of $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$. □

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 $\mathbf{A}$, and let $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$ be the lift of $\mathbf{A}$ obtained by marking C by a new unary symbol M_{C}.

Then, as every conservative lift of $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$ is a conservative lift of $\mathbf{A}$, it follows that every universal cluster of $\mathbf{A}$ is a universal cluster of $\mathrm{L}_{\mathsf{C}}(\mathbf{A})$.

The universal clusters of $\mathbf{A}$ 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 $\mathbf{A}$.

Lemma 11. *Let* $\mathbf{A}$ *be a local convergent sequence and let* ϕ *be a local formula with single free variable* x_1*.*

Then the following conditions are equivalent:

(1) $\partial_{\mathbf{A}}\phi(\mathbf{A}) \approx 0$*;*
(2) $\phi(\mathbf{A})$ *is a cluster of* $\mathbf{A}$*;*
(3) $\phi(\mathbf{A})$ *is a universal cluster of* $\mathbf{A}$*.*

Proof. If $\phi(\mathbf{A})$ is a (universal) cluster of $\mathbf{A}$ then $\partial_{\mathbf{A}}\phi(\mathbf{A}) \approx 0$ (by definition of a cluster).

Conversely, if $\partial_{\mathbf{A}}\phi(\mathbf{A}) \approx 0$ then either $\phi(\mathbf{A})$ is negligible (thus $\phi(\mathbf{A})$ is a cluster) or $\mathbf{A}[\phi(\mathbf{A})]$ is local convergent: for every local formula ψ with free variables $x_1, \dots, x_p$, denoting $\hat{\psi}$ the formula obtained by replacing terms $(\exists y)\theta$ by $(\exists y)(\phi(y) \wedge \theta)$ and terms $(\forall y)\theta$ by $(\forall y)(\phi(y) \to \theta)$ and denoting $\widetilde{\psi}$ the formula $\hat{\psi} \wedge \bigwedge_{i=1}^{p} \phi(x_i)$ we get $\widetilde{\psi}(\mathbf{A} \setminus \partial_{\mathbf{A}}\phi(\mathbf{A})) = \psi(\mathbf{A}[\phi(\mathbf{A})])$ hence

$$\langle \psi, \mathbf{A}_n[\phi(\mathbf{A}_n)] \rangle = \langle \widetilde{\psi}, \mathbf{A}_n \rangle + o(1).$$

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

Lemma 12. *Let* $\mathbf{A}$ *be a local convergent sequence. Then the equivalence classes of universal clusters of* $\mathbf{A}$ *form a Boolean algebra.*

Proof. Let X, Y be universal clusters of $\mathbf{A}$, and let $\mathrm{L}(\mathbf{A})$ be a local convergent conservative lift of $\mathbf{A}$. Then the sequence $\mathrm{L}(\mathrm{L}_{\mathsf{X}}(\mathrm{L}_{\mathsf{Y}}(\mathbf{A})))$ is local convergent. If follows, by considering formulas $\neg M_{\mathsf{X}}$, $M_{\mathsf{X}} \vee M_{\mathsf{Y}}$, and $M_{\mathsf{X}} \wedge M_{\mathsf{Y}}$ that $\mathsf{A} \setminus \mathsf{X}$, $\mathsf{X} \cup \mathsf{Y}$ and $\mathsf{X} \cap \mathsf{Y}$ are clusters of $\mathrm{L}(\mathrm{L}_{\mathsf{X}}(\mathrm{L}_{\mathsf{Y}}(\mathbf{A})))$ hence of $\mathrm{L}(\mathbf{A})$. It follows that $\mathsf{A} \setminus \mathsf{X}$, $\mathsf{X} \cup \mathsf{Y}$ and $\mathsf{X} \cap \mathsf{Y}$ are universal clusters of $\mathbf{A}$. □

5.3. **Pre-Clusters.**

Definition 10. A sequence $\mathsf{X} \not\approx \mathsf{0}$ is a *pre-cluster* of $\mathbf{A}$ if $\mathsf{X} \approx \mathsf{0}$ or if it holds

(1) the sequence $\mathbf{A}[\mathsf{X}]$ is local-convergent;
(2) the limit $\lim \nu_{\mathbf{A}}(\mathsf{X})$ and is strictly positive;
(3) for every integer d it holds

$$\limsup \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\mathsf{X}) \setminus \mathsf{X}) = 0.$$

The definition of pre-clusters of $\mathbf{A}$ is consistent with the notion of equivalence of sequence of subsets:

Lemma 13. *Let* X *be a pre-cluster of* $\mathbf{A}$ *and let* $\mathsf{Y} \approx \mathsf{X}$ *in* $\mathbf{A}$*. Then* Y *is a pre-cluster of* $\mathbf{A}$*.*

Proof. That $\mathbf{A}[\mathsf{Y}]$ is local-convergent follows from Lemma 3. Also, it is immediate that $\lim \nu_{\mathbf{A}}(\mathsf{Y})$ exists and that $\lim \nu_{\mathbf{A}}(\mathsf{Y}) = \lim \nu_{\mathbf{A}}(\mathsf{X})$.

Let $\mathsf{Z} = \mathsf{X} \,\Delta\, \mathsf{Y}$. By assumption, Z is negligible in $\mathbf{A}$.

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

$$\begin{aligned} \mathrm{N}^d_{\mathbf{A}}(\mathsf{Y}) \setminus \mathsf{Y} &\subseteq (\mathrm{N}^d_{\mathbf{A}}(\mathsf{X}) \cup \mathrm{N}^d_{\mathbf{A}}(\mathsf{Z})) \setminus (\mathsf{X} \setminus \mathsf{Z}) \\ &\subseteq (\mathrm{N}^d_{\mathbf{A}}(\mathsf{X}) \setminus \mathsf{X}) \cup \mathrm{N}^d_{\mathbf{A}}(\mathsf{Z}) \end{aligned}$$

It follows that $\limsup \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\mathsf{Y}) \setminus \mathsf{Y}) = 0$ hence Y is a pre-cluster. □

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

Proof. This follows from the fact that $\mathrm{N}^d_{\mathbf{A}}(\mathsf{X}) \setminus \mathsf{X} \subseteq \mathrm{N}^d_{\mathbf{A}}(\partial_{\mathbf{A}_n} \mathsf{X})$. □

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

Definition 11. Let X be a pre-cluster of a local-convergent sequence $\mathbf{A}$.

The *wrapping* of X in $\mathbf{A}$ 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 $\nu_{\mathbf{A}_{n'}}(\mathrm{N}^{2d+1}_{\mathbf{A}_{n'}}(X_{n'}) \setminus X_{n'}) < 1/d$. Then we define $W_n = \mathrm{N}^{D(n)}_{\mathbf{A}_n}(X_n)$.

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

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

$$\limsup \nu_{\mathbf{A}}(\mathsf{W} \setminus \mathsf{X}) = 0.$$

Proof. For every $d \in \mathbb{N}$ there exists $T(d)$ such that for every $n \geq T(d)$ it holds $\nu_{\mathbf{A}_n}(\mathrm{N}^{2d_1}_{\mathbf{A}_n}(X_n) \setminus X_n) < 1/d$. For $T(d) \leq n < T(d+1)$ we have $W_n = \mathrm{N}^d_{\mathbf{A}_n}(X_n)$. Thus, for every $d \in \mathbb{N}$ and every $T(d') \leq n < T(d'+1)$ (with $d' \geq d$) it holds

$$\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} W_n)) \leq \nu_{\mathbf{A}_n}(\mathrm{N}^{2d'+1}_{\mathbf{A}_n}(X_n) \setminus X_n) < 1/d'.$$

Thus $\partial_{\mathbf{A}} \mathsf{W}$ is negligible in $\mathbf{A}$ hence W is a cluster of $\mathbf{A}$.

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

Assume that a cluster Y of $\mathbf{A}$ as the same properties. Then

$$\limsup \nu_{\mathbf{A}}(\mathsf{W} \,\Delta\, \mathsf{Y}) \leq \limsup \nu_{\mathbf{A}}(\mathsf{W} \setminus \mathsf{X}) + \limsup \nu_{\mathbf{A}}(\mathsf{Y} \setminus \mathsf{X}) = 0.$$

Hence, according to Lemma 4, W and Y are equivalent in $\mathbf{A}$. □

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

Definition 12. A structure $\mathbf{A}$ is (d,ϵ,δ)-*expanding* if, for every $X \subset A$ it holds

$$\epsilon < \nu_{\mathbf{A}}(X) < 1-\epsilon \quad \Longrightarrow \quad \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(X)) > (1+\delta)\nu_{\mathbf{A}}(X),$$

that is

$$\inf\left\{ \frac{\nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}[X] \setminus X)}{\nu_{\mathbf{A}}(X)} : \quad \epsilon < \nu_{\mathbf{A}}(X) < 1-\epsilon \right\} > \delta.$$

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

$$h_{\mathrm{out}} = \inf\left\{ \frac{|\mathrm{N}_{\mathbf{A}}[X] \setminus X|}{|X|} : \quad 0 < \frac{|X|}{|A|} < 1/2 \right\}.$$

Lemma 16. *Let $0 < \epsilon < 1/6$ and let $\mathbf{A}$ be a (d,ϵ,δ)-expanding structure. Then there exists a subset $Y \subseteq A$ of measure $\nu_{\mathbf{A}}(Y) \leq \epsilon$ such that, denoting $\mathbf{A}' = \mathbf{A} - Y$, it holds*

$$\forall X \subseteq A' \qquad \nu_{\mathbf{A}'}(X) \leq 1/2 \implies \nu_{\mathbf{A}'}(\mathrm{N}^d_{\mathbf{A}'}(X) \setminus X) \geq \delta \nu_{\mathbf{A}'}(X).$$

Proof. Let $Y \subseteq A$ be maximal (for inclusion) with the property that $\nu_{\mathbf{A}}(Y) \leq 1-2\epsilon$ and $\nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(Y) \setminus Y) < \delta \nu_{\mathbf{A}}(Y)$. First note that $\nu_{\mathbf{A}}(Y) \leq \epsilon$ as $\mathbf{A}$ is (d,ϵ,δ)-expanding. Let $\mathbf{A}' = \mathbf{A} - Y$.

Assume for contradiction that there exists $Z \subset A'$ is such that $\nu_{\mathbf{A}'}(Z) \leq 1/2$ and $\nu_{\mathbf{A}'}(\mathrm{N}^d_{\mathbf{A}'}(Z) \setminus Z) < \delta \nu_{\mathbf{A}'}(Z)$. (Note that $\nu_{\mathbf{A}}(Z) \leq n u_{\mathbf{A}'}(Z) \leq 1/2$.) As $\nu_{\mathbf{A}'}$ is proportional to $\nu_{\mathbf{A}}$ it also holds $\nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}'}(Z) \setminus Z) < \delta \nu_{\mathbf{A}}(Z)$. Moreover it obviously holds $\mathrm{N}^d_{\mathbf{A}}(Y \cup Z) \subseteq \mathrm{N}^d_{\mathbf{A}}(Y) \cup \mathrm{N}^d_{\mathbf{A}'}(Z)$ thus

$$\begin{aligned} \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(Y \cup Z)) &\leq \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(Y)) + \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}'}(Z)) \\ &< (1+\delta)(\nu_{\mathbf{A}}(Y) + \nu_{\mathbf{A}}(Z)) = (1+\delta)(\nu_{\mathbf{A}}(Y \cup Z)) \end{aligned}$$

Hence $\nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(Y \cup Z) \setminus (Y \cup Z)) < \delta \nu_{\mathbf{A}}(Y \cup Z)$. However $\nu_{\mathbf{A}}(Y \cup Z) = \nu_{\mathbf{A}}(Y) + \nu_{\mathbf{A}}(Z) \leq \epsilon + 1/2 < 1 - 2\epsilon$, what contradicts the maximality of Y. □

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,ϵ,δ)-expansion means that the dth power of the graph (after deletion of a subset of vertices of measure at most ϵ) has magnification at least δ. In the very special (but standard) case of graphs with maximum degree at most Δ we recover the standard definition of expansion:

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

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 Δ then for every subset X of vertices and for every integer $k \geq 1$ it holds $|\mathrm{N}^{k+1}_G(X) \setminus \mathrm{N}^k_G(X)| \leq (\Delta-1)|\mathrm{N}^k_G(X) \setminus \mathrm{N}^{k-1}_G(X)|$, where we define $\mathrm{N}^0_G(X) = X$. Hence $|\mathrm{N}^d_G(X) \setminus X| \leq (1 + \cdots + (\Delta-1)^{d-1})|\mathrm{N}_G(X) \setminus X|$. □

Definition 13. A local-convergent sequence $\mathbf{A}$ is *expanding* if, for every $\epsilon > 0$ there exist $d, t \in \mathbb{N}$ and $\delta > 0$ such that every $\mathbf{A}_n$ with $n \geq t$ is (d,ϵ,δ)-expanding.

We have the following equivalent formulations of this concept:

Lemma 18. *Let $\mathsf{X} \not\approx \mathsf{0}$ be a cluster of a local convergent sequence $\mathbf{A}$. The following conditions are equivalent:*

(1) *the sequence* $\mathbf{A}[\mathsf{X}]$ *is expanding;*
(2) *for every* $\epsilon > 0$ *there exists* $d, t \in \mathbb{N}$ *such that for every* $\mathsf{Z} \subseteq \mathsf{X}$ *with* $\nu_{\mathbf{A}_n}(Z_n) > \epsilon\nu_{\mathbf{A}_n}(X_n)$ *and every* $n \geq t$ *it holds*

$$\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(Z_n)) > (1-\epsilon)\nu_{\mathbf{A}_n}(X_n);$$

(3) *the sequence* X *is a strongly atomic cluster of* $\mathbf{A}$*;*
(4) *for every* $\epsilon > 0$ *there exists no* $\mathsf{Y} \subseteq \mathsf{X}$ *such that* $\partial_{\mathbf{A}}\mathsf{Y} \approx 0$ *and*

$$\epsilon < \liminf \nu_{\mathbf{A}}(\mathsf{Y}) < \lim \nu_{\mathbf{A}}(\mathsf{X}) - \epsilon.$$

Proof. First assume that $\mathbf{A}[\mathsf{X}]$ is expanding, and assume for contradiction that X is not a strongly atomic cluster of $\mathbf{A}$. Then there exists some increasing function $f : \mathbb{N} \to \mathbb{N}$ such that Y_f is a non-trivial cluster of $\mathbf{A}_f$, $\mathsf{Y}_f \subseteq \mathsf{X}_f$ and $\mathsf{Y}_f \not\approx \mathsf{X}_f$. Then $\alpha = \lim \nu_{\mathbf{A}_f}(\mathsf{Y}_f)/\nu_{\mathbf{A}_f}(\mathsf{X}_f)$ is bounded away from 0 and 1. Thus there exists $\delta > 0$ and $d \in \mathbb{N}$ such that

$$\liminf \frac{\nu_{\mathbf{A}_f}(\mathrm{N}^d_{\mathbf{A}_f[\mathsf{X}_f]}(\mathsf{Y}_f))}{\nu_{\mathbf{A}_f}(\mathsf{Y}_f)} > 1 + \delta,$$

what contradicts the property that Y_f is a cluster.

Now assume that X is a strongly atomic cluster of $\mathbf{A}$ and assume for contradiction that $\mathbf{A}[\mathsf{X}]$ is not expanding. Then there exists $\epsilon > 0$ such that for every $d \in \mathbb{N}$ it holds

$$\liminf_{n\to\infty} \inf_{Y_n} \frac{\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n[X_n]}(Y_n))}{\nu_{\mathbf{A}_n}(Y_n)} = 1,$$

where infimum is on subsets $Y_n \subset X_n$ with $\epsilon < \nu_{\mathbf{A}_n}(Y_n)/\nu_{\mathbf{A}_n}(X_n) < 1 - \epsilon$. We inductively construct an increasing function $f : \mathbb{N} \to \mathbb{N}$ and subsets $Y_{f(n)} \subset X_{f(n)}$ as follows: $f(1)$ is the minimum integer n such that there exists $Y_n \subset X_n$ with $\epsilon < \nu_{\mathbf{A}_n}(Y_n)/\nu_{\mathbf{A}_n}(X_n) < 1 - \epsilon$ and $\nu_{\mathbf{A}_n}(\mathrm{N}_{\mathbf{A}_n[X_n]}(Y_n)) < 2\nu_{\mathbf{A}_n}(Y_n)$ and (for $d \geq 1$) $f(d+1)$ is the minimum integer $n > f(d)$ such that there exists $Y_n \subset X_n$ with $\epsilon < \nu_{\mathbf{A}_n}(Y_n)/\nu_{\mathbf{A}_n}(X_n) < 1 - \epsilon$ and $\nu_{\mathbf{A}_n}(\mathrm{N}^{d+1}_{\mathbf{A}_n[X_n]}(Y_n)) < \frac{d+2}{d+1}\nu_{\mathbf{A}_n}(Y_n)$. It is easily checked that $(Y_{f(n)})$ is such that for every integer d it holds

$$\limsup \nu_{\mathbf{A}_f}(\mathrm{N}^d_{\mathbf{A}_f}(\mathsf{Y}_f) \setminus \mathsf{Y}_f) = 0.$$

We can further consider a subsequence $\mathsf{Y}_{f\circ g}$ of Y_f such that $\mathbf{A}_{f\circ g}[\mathsf{Y}_{f\circ g}]$ is local convergent and $\nu_{\mathbf{A}_{f\circ g}}(\mathsf{Y}_{f\circ g})$ converges. It follows that $\mathsf{Y}_{f\circ g}$ is a pre-cluster. Let ($\hat{\mathsf{Y}}_{f\circ g}$ be the wrapping of $\mathsf{Y}_{f\circ g}$ in $\mathbf{A}_{f\circ g}$). Then $\hat{\mathsf{Y}}_{f\circ g}$ is a cluster, $\hat{\mathsf{Y}}_{f\circ g} \preceq \mathsf{X}_{f\circ g}$ and $\hat{\mathsf{Y}}_{f\circ g} \not\approx \mathsf{X}_{f\circ g}$, what contradicts the assumption that X is a strongly atomic cluster. □

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

Definition 14. A local-convergent sequence $\mathbf{A}$ is *non-dispersive* if, for every $\epsilon > 0$ there exists $d \in \mathbb{N}$ such that

$$\liminf_{n\to\infty} \sup_{v_n \in A_n} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(v_n)) > 1 - \epsilon.$$

In other words, a sequence $\mathbf{A}$ is non-dispersive if, for every $\epsilon > 0$, ϵ-almost all elements of $\mathbf{A}_n$ are included in some ball of radius at most d, for some fixed d.

Definition 15. A non-trivial cluster X of $\mathbf{A}$ is a *globular cluster* of $\mathbf{A}$ if $\mathbf{A}[\mathsf{X}]$ 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 16. A cluster X of $\mathbf{A}$ is *residual* if for every $d \in \mathbb{N}$ it holds

$$\limsup_{n\to\infty} \sup_{v_n \in A_n} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(v_n)) = 0.$$

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

$$\liminf h_{\mathrm{out}}(G_n) \geq \alpha.$$

Lemma 19. *Let $\mathbf{G}$ be a sequence of graphs with maximum degree at most Δ and let $\mathsf{C} \not\approx \mathsf{0}$ be a cluster of $\mathbf{G}$. The following are equivalent:*

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

Proof. This is a direct consequence of Lemma 17. □

Lemma 20. *Let X be an expanding cluster of $\mathbf{A}$, and let Y be any cluster of $\mathbf{A}$.*
Then any convergent subsequence of $\left(\frac{\nu_{\mathbf{A}}(\mathsf{X}\cap\mathsf{Y})}{\nu_{\mathbf{A}}(\mathsf{X})}\right)$ has limit either 0 or 1.

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

According to Fact 1, for every integer $d \in \mathbb{N}$ it holds

$$\mathrm{N}^d_{\mathbf{A}_{f(n)}}(Z_{f(n)}) \subseteq \mathrm{N}^{d+1}_{\mathbf{A}_{f(n)}}(X_{f(n)}) \cup \mathrm{N}^{d+1}_{\mathbf{A}_{f(n)}}(Y_{f(n)}).$$

It follows that $\partial_{\mathbf{A}_f} \mathsf{Z}_f$ is negligible in $\mathbf{A}_f$.

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

Lemma 21. *Let X and Y be expanding clusters of $\mathbf{A}$.*
Then

- *either $\mathsf{X} \cap \mathsf{Y} \approx \mathsf{0}$ (i.e. X and Y are* essentially disjoint*);*
- *or $\mathsf{X} \between \mathsf{Y}$ and then every Z with $Z_n \in \{X_n, Y_n\}$ is an expanding cluster of $\mathbf{A}$.*

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

Otherwise, there exists, according to Lemma 20, an increasing function $f : \mathbb{N} \to \mathbb{N}$ such that $\mathsf{X}_f \approx \mathsf{Y}_f$. It follows that $\mathbf{A}_f[\mathsf{X}_f]$ and $\mathbf{A}_f[\mathsf{Y}_f]$ (thus $\mathbf{A}[\mathsf{X}]$ and $\mathbf{A}[\mathsf{Y}]$) have the same local limit. Let Z be such that $Z_n \in \{X_n, Y_n\}$. Then $\mathbf{A}[\mathsf{Z}]$ is local-convergent, $\partial_{\mathbf{A}}\mathsf{Z}$ is negligible in $\mathbf{A}$, and $\lim \nu_{\mathbf{A}}(\mathsf{Z})$ exists (and $\lim \nu_{\mathbf{A}}(\mathsf{Z}) = \lim \nu_{\mathbf{A}}(\mathsf{X}) = \lim \nu_{\mathbf{A}}(\mathsf{Y})$). It follows that Z is a non-trivial cluster. Thus X and Y are interweaving (i.e. $\mathsf{X} \between \mathsf{Y}$). 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. □

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

Example 2. Consider a local-convergent sequence $\mathbf{E}$ of connected d-regular high-girth expanders with $|E_n| = cn(1 + o(1))$ (and uniform probability measure), for some constant $c > 1$. Let $\mathbf{A}_n$ be defined as three copies of $\mathbf{E}_n$ if n is odd, and the union of $\mathbf{E}_n$ and $\mathbf{E}_{2n}$ if n is even. Selecting a copy of $\mathbf{E}_n$ into each $\mathbf{A}_n$ leads to uncountably many pairwise intersecting non-equivalent expanding clusters. However,

no two essentially disjoint expanding clusters exist in $\mathbf{A}$. Note that we could have made $\mathbf{A}_n$ connected by adding paths of length $\sqrt{n}$ linking the connected components without changing the conclusion.

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 17. Let $\mathbf{A}$ be a local-convergent sequence of σ-structures. A lifted sequence $\mathrm{L}(\mathbf{A})$ of $\mathbf{A}$ obtained by extending the signature σ into σ^+ by adding countably many unary relations $M_1, M_2, \dots$ is a *clustering* of $\mathbf{A}$ if, denoting

$$\mathsf{S} = \mathsf{A} \setminus \bigcup M_i(\mathbf{A})$$

the following conditions hold:

(1) The sequence $\mathrm{L}(\mathbf{A})$ is local convergent;
(2) The sequence S is negligible and $\bigcup_i \partial_{\mathbf{A}} M_i(\mathbf{A}) \subseteq \mathsf{S}$;
(3) For every $n \in \mathbb{N}$, the non empty sets among S_n, $M_1(\mathbf{A}_n), M_2(\mathbf{A}_n), \dots$ form a partition of A_n;
(4) The partition induced by the M_i's is stable in the sense that

$$\sum_i \lim \langle M_i, \mathbf{A}\rangle = \lim \sum_i \langle M_i, \mathbf{A}\rangle = 1.$$

Definition 18. We say that two clusters C_1 and C_2 are

- *weakly disjoint* if $\mathsf{C}_1 \Delta \mathsf{C}_2 \approx \mathsf{0}$;
- *disjoint* if $\mathsf{C}_1 \Delta \mathsf{C}_2 = \mathsf{0}$;
- *strongly disjoint* if $(\mathrm{N}_{\mathbf{A}}(\mathsf{C}_1) \cap \mathsf{C}_2) \cup (\mathsf{C}_1 \cap \mathrm{N}_{\mathbf{A}}(\mathsf{C}_2)) = \mathsf{0}$.

Remark 2. Conditions (1) and (2) imply that each sequence $M_i(\mathbf{A})$ is a cluster of $\mathbf{A}$ hence a clustering of $\mathbf{A}$ induce a "partition" into countably many disjoint clusters, and that the clusters defined by the marks M_i are pairwise strongly disjoint.

A simple idea to construct a clustering of a local convergent sequence $\mathbf{A}$ is as follows: assume $\mathbf{A}$ has a cluster $\mathsf{X}_1 \not\approx \mathsf{0}$. Then let $\mathbf{A}_1$ be the lift of $\mathbf{A}$ with X_1 marked M_1. Then look for a cluster $\mathsf{X}_2 \not\approx \mathsf{0}$ of $\mathbf{A}_1$ disjoint from M_1 and mark it M_2, thus obtaining $\mathbf{A}_2$. 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 $\bigcup_i \partial_{\mathbf{A}} M_i(\mathbf{A})$ needs not to be negligible;
- The partition is essentially not unique (and it is not clear which clusters of $\mathbf{A}$ may appear simultaneously in the partition).

The first point is exemplified by the fact that we do not have the converse of Remark 2 does not holds in general: partitioning into disjoint clusters do not define a clustering in general.

For instance, consider the following sequence of star forests.

Example 3. Consider the sequence $\mathbf{G}$ where $\mathbf{G}_n$ is the union of 2^n stars $\mathbf{H}_{n,1}, \dots, \mathbf{H}_{n,2^n}$, where the i-th star $\mathbf{H}_{n,i}$ has order $2^{2^n}(2^{-i} + 2^{-n})/2$. Let C^i be the sequence such that C^i_n is the vertex set $H_{n,i}$ of the ith biggest connected component of $\mathbf{G}_n$ (or the empty subset if $i > 2^n$). It is easily checked that each C^i is a cluster and that for each n the (non-empty) subsets C^i_n form a partition of G_n. Assume that we mark each C^i_n 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 $\mathsf{C}^1, \dots, \mathsf{C}^i, \dots$ of a local convergent sequence $\mathbf{A}$. For each $i \in \mathbb{N}$ we define

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

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* $\mathsf{S} \supseteq \bigcup_i \partial_{\mathbf{A}} \mathsf{C}^i$ *and that it holds*

$$\sum_i \lim \nu_{\mathbf{A}}(\mathsf{C}^i) = \lim \sum_i \nu_{\mathbf{A}}(\mathsf{C}^i). \tag{1}$$

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

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

$$\begin{aligned}
\left| \left(\sum_{i \geq 1} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \phi^i, \mathbf{A}_n[C_n^i] \rangle \right) - \left(\sum_{i \geq 1} \lambda_i^p \lim \langle \phi^i, \mathbf{A}[\mathsf{C}^i] \rangle \right) \right| \leq \\
\sum_{i \geq 1} |\nu_{\mathbf{A}_n}(C_n^i)^p - \lambda_i^p| + \sum_{i \geq 1} \lambda_i^p \left| \lim \langle \phi^i, \mathbf{A}[\mathsf{C}^i] \rangle - \langle \phi^i, \mathbf{A}_n[C_n^i] \rangle \right| \\
\| \nu_{\mathbf{A}}(\mathsf{C}^i) - (\lambda_i)_{i \in \mathbb{N}} \|_p + \sum_{i=1}^{N} \left| \lim \langle \phi^i, \mathbf{A}[\mathsf{C}^i] \rangle - \langle \phi^i, \mathbf{A}_n[C_n^i] \rangle \right| + \sum_{i > N} \lambda_i^p.
\end{aligned}$$

It follows that

$$\lim_{n \to \infty} \sum_{i \geq 1} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \phi^i, \mathbf{A}_n[C_n^i] \rangle = \sum_{i \geq 1} \lambda_i^p \lim \langle \phi^i, \mathbf{A}[\mathsf{C}^i] \rangle. \tag{2}$$

Let ψ be a strongly r-local formula with p free variables in the language of $\sigma^+ = \sigma \cup \{M_0, M_1, M_2, \dots\}$. For ζ non negative integer, let ψ^ζ be the formula obtained from ψ by replacing each term $M_i(t)$ by `true` if $i = \zeta$ and *false* otherwise. According to Lemma 2 it holds

$$\begin{aligned}
\langle \psi, \mathrm{L}(\mathbf{A}_n) \rangle &= \nu_{\mathbf{A}_n}(R_n)^p \langle \psi^0, \mathbf{A}_n[R_n] \rangle + \sum_{i \geq 1} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \psi^i, \mathbf{A}_n[C_n^i] \rangle, \\
\langle \psi^0, \mathbf{A}_n \rangle &= \nu_{\mathbf{A}_n}(R_n)^p \langle \psi^0, \mathbf{A}_n[R_n] \rangle + \sum_{i \geq 1} \nu_{\mathbf{A}_n}(C_n^i)^p \langle \psi^0, \mathbf{A}_n[C_n^i] \rangle.
\end{aligned}$$

Thus, according to (2) it holds

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

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

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

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

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

Hence $\mathrm{L}(\mathbf{A})$ is a clustering of $\mathbf{A}$. □

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

Definition 19. A *clip* us 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 every integers $n \leq n'$ it holds

$$\sum_{i=1}^{F(n)} \left|\nu_{\mathbf{A}_{n'}}(C^i_{n'}) - \lambda_i\right| \leq \sum_{i>F(n)} \lambda_i. \tag{3}$$

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

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

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

$$\sum_{i=1}^{G(n)} \left|\nu_{\mathbf{A}_{n'}}(C^i_{n'}) - \lambda_i\right| \leq \sum_{i=1}^{F(n)} \left|\nu_{\mathbf{A}_{n'}}(C^i_{n'}) - \lambda_i\right| \leq \sum_{i>F(n)} \lambda_i \leq \sum_{i>G(n)} \lambda_i.$$

Lemma 23. *Let F be a clip. Then*

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

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

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

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

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

□

Given a clip F, we define R by

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

As for every integers i and d it holds

$$\lim \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\partial_{\mathbf{A}}\mathsf{C}^{\mathsf{i}})) = 0$$

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

$$\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} C^i_n)) \leq \frac{2^{-i}}{d}.$$

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*

$$S_n = \bigcup_{i=1}^{G(n)} \partial_{\mathbf{A}_n} C_n^i$$

is negligible.

Proof. Let $d \in \mathbb{N}$. For sufficiently large n it holds

$$\begin{aligned}\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(S_n)) &\leq \nu_{\mathbf{A}_n}(\mathrm{N}^{G(n)}_{\mathbf{A}_n}(S_n)) \\ &\leq \sum_{i=1}^{G(n)} \nu_{\mathbf{A}_n}(\mathrm{N}^{G(n)}_{\mathbf{A}_n}(\partial C_n^i)) \\ &\leq \sum_{i=1}^{G(n)} \frac{2^{-i}}{G(n)} < \frac{1}{G(n)}.\end{aligned}$$

Hence

$$\lim \nu_{\mathbf{A}}(\mathrm{N}^d_{\mathbf{A}}(\mathsf{S})) = 0,$$

that is: S is negligible. □

Define the subset sequences D^i by

$$D_n^i = \begin{cases} C_n^i & \text{if } n \geq G(i) \\ \emptyset & \text{otherwise} \end{cases}$$

and let R be defined by $R_n = A_n \setminus \bigcup_{i\geq 1} D_n^i$.

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

Proof. Note that

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

In particular, $\partial\mathsf{R}$ is negligible. According to Lemma 23, we have

$$\lim_{n\to\infty} \sum_{i=1}^{G(n)} \nu_{\mathbf{A}_n}(D_n^i) = 1 - \lambda_0,$$

thus $\lim \nu_{\mathbf{A}}(\mathsf{R}) = \lambda_0$. Consider a strongly r-local formula ϕ with p free variable. For every $\epsilon > 0$ there exists n_0 such that for every $n \geq n_0$ it holds $\nu_{\mathbf{A}_n}(\mathrm{N}^r_{\mathbf{A}_n}(S_n)) < \epsilon$. It follows that

$$\left| \langle \phi, \mathbf{A}_n \rangle - \nu_{\mathbf{A}_n}(R_n)^p \langle \phi, \mathbf{A}_n[R_n] \rangle - \sum_{i\geq 1} \nu_{\mathbf{A}_n}(D_n^i)^p \langle \phi, \mathbf{A}_n[D_n^i] \rangle \right| < 3p\epsilon.$$

Thus, if $\lambda_0 > 0$ we have

$$\begin{aligned}\lim \langle \phi, \mathbf{A}[\mathsf{R}] \rangle &= \frac{1}{\lambda_0^p} \lim_{n\to\infty} \Big(\langle \phi, \mathbf{A}_n \rangle - \sum_{i\geq 1} \lambda_i^p \langle \phi, \mathbf{A}_n[D_n^i] \rangle \Big) \\ &= \frac{1}{\lambda_0^p} \Big(\lim \langle \phi, \mathbf{A} \rangle - \sum_{i\geq 1} \lambda_i^p \lim \langle \phi, \mathbf{A}[\mathsf{D}^i] \rangle \Big).\end{aligned}$$

(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 λ_0 and R is negligible, or $\lambda_0 > 0$ and R is a cluster. □

Lemma 26 (Cluster Comb Lemma)**.** *Let* $\mathbf{A}$ *be a local convergent sequence of* σ*-structures, and let* $\mathsf{C}^1, \dots, \mathsf{C}^i, \dots$ *be countably many strongly disjoint clusters of* $\mathbf{A}$*.*

Let σ^+ *be the signature* σ *augmented by unary relations* M_i *(*$i \in \{0, 1, 2, \dots\}$*). Then there exist a clustering* $\mathbf{A}^+$ *of* $\mathbf{A}$ *with the property that for* $i = 1, \dots,$ *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*

$$M_i(\mathbf{A}_n) = \begin{cases} C_n^i & \text{if } n \geq G(i) \\ \emptyset & \text{otherwise} \end{cases} \tag{4}$$

(In particular $M_i(\mathbf{A}) \approx \mathsf{C}^i$*).*

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

Remark that we have the property that $\mathbf{A} \approx \mathbf{A} - \mathsf{S}$, which is the disjoint union of $\mathbf{A}[R]$ and all the $\mathbf{A}[\mathsf{D}^i]$. 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 σ^+ extended by these marks, and let $\mathbf{A}^+$ be the sequence of marked structures. Let ϕ be an r-local strongly local formula with p free variables. Then

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

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

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

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

Remark: if one only assumes that the clusters C^i are almost disjoint (meaning $\mathsf{C}^i \Delta \mathsf{C}^j$ negligible if $i \neq j$) then we get the same conclusion, except that the second item is weakened to $\mathsf{D}^i \approx \mathsf{C}^i$. The idea is to define the clusters $\mathsf{Z}^i = \mathsf{C}^i \setminus \bigcup_{j<i} \mathrm{N}_{\mathbf{A}}(\mathsf{C}^j)$ 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 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. 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 $\mathcal{B}$ be the Lindenbaum–Tarski algebra defined by $\mathrm{FO}^{\mathrm{local}}(\sigma)$ and let S_σ be the Stone dual of $\mathcal{B}$, which is a Polish space, whose topology is generated by its clopen subsets. Recall that the duality of $\mathcal{B}$ and S_σ is expressed by the existence of a mapping K from $\mathrm{FO}^{\mathrm{local}}(\sigma)$ to the family of all the clopen subsets of S such that $K(\phi \vee \psi) = K(\phi) \cup K(\psi)$, $K(\phi \wedge \psi) = K(\phi) \cap K(\psi)$, $K(\neg\phi) = S \setminus K(\phi)$, and $K(\phi) = K(\psi)$ if and only if ϕ and ψ are logically equivalent. For a local formula ϕ, we further denote by $k(\phi)$ the indicator function of $K(\phi)$, which is obviously continuous on S. Note that the σ-algebra of Borel subsets of S_σ turns S_σ into a standard Borel space.

The following representation theorem has been proved in [13] (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.

Theorem 3. *For every finite structure $\mathbf{A}$ there is a unique probability measure $\mu_{\mathbf{A}}$ on S_σ such that for every finite σ-structure $\mathbf{A}$ and every local formula ϕ it holds*

$$\langle \phi, \mathbf{A} \rangle = \int_{S_\sigma} k(\phi)\, \mathrm{d}\mu_{\mathbf{A}}.$$

Moreover, for every two finite structures $\mathbf{A}$ and $\mathbf{B}$, it holds $\mu_{\mathbf{A}} = \mu_{\mathbf{B}}$ if and only if the structures obtained from $\mathbf{A}$ and $\mathbf{B}$ 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 $\mu_{\mathbf{A}}$ for finite $\mathbf{A}$ (with respect to weak topology).

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

$$\int_{S_\sigma} k(\phi)\, \mathrm{d}\mu_{\lim \mathbf{A}} = \langle \phi, \lim \mathbf{A} \rangle.$$

Recall that a bounded sequence of positive finite measures μ_n on a metric space S *converges weakly* to the finite positive measure μ if for any bounded continuous function $f : S \to \mathbb{R}$ it holds $\int f\, \mathrm{d}\mu_n \to \int f\, \mathrm{d}\mu$. 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 $\mathbf{A}$ it holds $\mu_{\mathbf{A}_n} \Rightarrow \mu_{\lim \mathbf{A}}$ and thus

$$\int_{S_\sigma} f\, \mathrm{d}\mu_{\lim \mathbf{A}} = \lim_{n\to\infty} \int_S f\, \mathrm{d}\mu_{\mathbf{A}_n}. \tag{5}$$

(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 $X_n \xrightarrow{\mathcal{D}} X$. 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 $F_X(x) = [X \leq x]$.

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 $\mathbf{A} = (\mathbf{A}_n)_{n\in\mathbb{N}}$ of finite structures is local convergent then so is the sequence $(F(\mathbf{A}_n))_{n\in\mathbb{N}}$ of the associated parameters. Of course, every continuous real function $f \in C(S_\sigma)$ defines a local continuous parameter $\mathbf{A} \mapsto \int_{S_\sigma} f \,\mathrm{d}\mu_{\mathbf{A}}$. But we shall 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 $\mathbf{A}$ be a σ-structure and let ϕ be a first order formula with free variables $x_1, \ldots, x_p$ (with $p \geq 1$). Denote by $\phi^v(\mathbf{A})$ the set

$$\phi^v(\mathbf{A}) = \Big\{(u_1, \ldots, u_{p-1}) \in A^{p-1} : \ \mathbf{A} \models (v, u_1, \ldots, u_{p-1})\Big\}.$$

The *local Stone pairing* of ϕ and $\mathbf{A}$ at v is

$$\begin{aligned}\langle \psi, \mathbf{A}\rangle_v &= \Pr(\mathbf{A} \models \psi(v, X_2, \ldots, X_p)) \\ &= \nu_{\mathbf{A}}^{\otimes(p-1)}(\phi^v(\mathbf{A}))\end{aligned}$$

Hence if $\nu_{\mathbf{A}}(\{v\}) \neq 0$ we get that the local Stone pairing of ϕ and $\mathbf{A}$ at v is nothing but the conditional probability $\Pr(\mathbf{A} \models \psi(X_1, X_2, \ldots, X_p) | X_1 = v)$.

In our setting, every finite structure is considered as a probability space and thus the local Stone pairing of a formula ϕ and finite structure $\mathbf{A}$ defines a random variable

$$\langle \phi, \mathbf{A}\rangle_\bullet : v \mapsto \langle \phi, \mathbf{A}\rangle_v.$$

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 $\mathbf{X}$ with values in $\mathbb{R}^k$ and law P, the *characteristic function* of $\mathbf{X}$ or P is

$$\gamma(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}\cdot\mathbf{X}}] = \int e^{i\mathbf{t}\cdot\mathbf{x}} \,\mathrm{d}P(\mathbf{x}) \qquad \text{for every } \mathbf{t} \in \mathbb{R}^k,$$

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

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

$$\gamma(\mathbf{t}) = \sum_{w_1\geq 0} \cdots \sum_{w_k \geq 0} \mathbb{E}[X_1^{w_1} \ldots X_k^{w_k}] \frac{i t_1^{w_1} \ldots t_k^{w_k}}{w_1! \ldots w_k!}.$$

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 $\gamma_n(\mathbf{t})$ converge for all $\mathbf{t}$ to some $\gamma(\mathbf{t})$, 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 γ 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*

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

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

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

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 $\mu \in \mathfrak{M}_\sigma$ and $\mathbf{t} \in \mathbb{R}^d$ define

$$\gamma(\mu, \mathbf{t}) = \sum_{w_1\geq 0} \cdots \sum_{w_d \geq 0} \left(\int_{S_\sigma} k(\psi_{\mathbf{w}})\,\mathrm{d}\mu \right) \prod_{j=1}^{d} \frac{(it_j)^{w_j}}{w_j!},$$

where $\psi_{(0,\ldots,0)}$ is true statement, and for $\mathbf{w} \neq (0,\ldots,0)$ we define

$$\psi_{\mathbf{w}} := \bigwedge_{i=1}^{d} \bigwedge_{j=1}^{w_i} \phi_i(x_1, x_{n_{i,j}+1}, \ldots, x_{n_{i,j}+p_i-1})$$

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 $\mu \in \mathfrak{M}_\sigma$, the mapping $\mathbf{t} \mapsto \gamma(\mu, \mathbf{t})$ is the characteristic function of a d-dimensional random variable $\mathbf{D}(\mu)$;*
2. *the mapping $\mu \mapsto \mathbf{D}(\mu)$ is continuous in the sense that if μ_n converges weakly to μ then $D(\mu_n)$ converges in distribution to $D(\mu)$, that is:*

$$\mu_n \Rightarrow \mu \qquad \Longrightarrow \qquad \mathbf{D}(\mu_n) \xrightarrow{\mathcal{D}} \mathbf{D}(\mu);$$

3. *for every finite structure $\mathbf{A}$ (with associated probability measure $\mu_{\mathbf{A}} \in \mathfrak{M}_\sigma$) the d-dimensional random variable*

$$\mathbf{D}_{\mathbf{A}} = (\langle \phi_1, \mathbf{A}\rangle_\bullet, \ldots, \langle \phi_d, \mathbf{A}\rangle_\bullet)$$

has the same distribution as $\mathbf{D}(\mu_{\mathbf{A}})$.

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

Let us prove (3). For any finite structure $\mathbf{A}$ and any vector $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{N}^d$, let $N = n_{d,w_d} + p_d - 1$. Then it holds

$$\begin{aligned}
\psi_{\mathbf{w}}(\mathbf{A}) &= \left\{ \mathbf{x} \in A^N : \ (\forall i \in [d]\ \forall j \in [w_i])\ (x_1, x_{n_{i,j}+1}, \ldots, x_{n_{i,j}+p_i-1}) \in \phi_i(\mathbf{A}) \right\} \\
&= \bigcup_{v\in A} \{v\} \times \left\{ \mathbf{x} \in A^{N-1} : \ (\forall i \in [d]\ \forall j \in [w_i])\ (x_{n_{i,j}}, \ldots, x_{n_{i,j}+p_i-2}) \in \phi_i^v(\mathbf{A}) \right\} \\
&= \bigcup_{v\in A} \{v\} \times \overbrace{\phi_1^v(\mathbf{A}) \times \cdots \times \phi_1^v(\mathbf{A})}^{w_1 \text{ times}} \times \cdots \times \overbrace{\phi_d^v(\mathbf{A}) \times \cdots \times \phi_d^v(\mathbf{A})}^{w_d \text{ times}}.
\end{aligned}$$

Thus

$$\begin{aligned}
\langle \psi_{\mathbf{w}}, \mathbf{A}\rangle &= \nu_{\mathbf{A}}^{\otimes N}(\psi_{\mathbf{w}}(\mathbf{A})) \\
&= \sum_{v\in A} \nu_{\mathbf{A}}(\{v\})\ \left(\nu_{\mathbf{A}}^{\otimes(p_1-1)}(\phi_1^v(\mathbf{A}))\right)^{w_1} \ldots \left(\nu_{\mathbf{A}}^{\otimes(p_d-1)}(\phi_d^v(\mathbf{A}))\right)^{w_d} \\
&= \mathbb{E}_v[\langle \phi_1, \mathbf{A}\rangle_v^{w_1} \ldots \langle \phi_d, \mathbf{A}\rangle_v^{w_d}].
\end{aligned}$$

If follows that the characteristic function $\gamma_{\mathbf{A}}$ of $\mathbf{D}_{\mathbf{A}}$ is equal to:

$$\gamma_{\mathbf{A}}(\mathbf{t}) = \mathbb{E}[e^{i\mathbf{t}\cdot\mathbf{D}_{\mathbf{A}}}] = \sum_{w_1\geq 0}\cdots\sum_{w_d\geq 0}\langle\psi_{\mathbf{w}},\mathbf{A}\rangle\prod_{j=1}^{d}\frac{(it_j)^{w_j}}{w_j!} = \gamma(\mu_{\mathbf{A}},\mathbf{t}).$$

(Note that as all the moments $\langle\psi_{\mathbf{w}},\mathbf{A}\rangle$ are bounded by 1 the above series converges for every (complex) vector $\mathbf{t}$.) As they have the same characteristic functions, the random variables $\mathbf{D}_{\mathbf{A}}$ and $\mathbf{D}(\mu_{\mathbf{A}})$ have the same distribution.

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

Remark 3 (for an interested reader). The formula defining $\psi_{\mathbf{w}}$ and the equality of $\langle\psi_{\mathbf{w}},\mathbf{A}\rangle$ and $\mathbb{E}_v[\langle\phi_1,\mathbf{A}\rangle_v^{w_1}\dots\langle\phi_d,\mathbf{A}\rangle_v^{w_d}]$ 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_2 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, $\langle(x_1\sim x_2)\wedge(x_1\sim x_3),G\rangle_v$ 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 $\langle(x_1\sim x_2)\wedge(x_1\sim x_3),G\rangle = (\langle x_1\sim x_2,G\rangle_v)^2]$. The same way, for every integer k, it holds

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

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 $\mu\in\mathfrak{M}_\sigma$ and let $(\phi_{\ell,1})_{\ell\in\mathbb{N}}$, …, $(\phi_{\ell,d})_{\ell\in\mathbb{N}}$ be sequences of local formulas (with $p_1,\dots,p_d$ free variables, respectively) such that for every integer $1\leq i\leq d$ it holds*

$$\phi_{1,i}\to\phi_{2,i}\to\dots\to\phi_{\ell,i}\to\dots$$

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

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

Then, as $\ell\to\infty$, the random variables $\mathbf{D}_\ell(\mu)$ converge in distribution to a random variable $\mathbf{D}_\infty(\mu)$, whose characteristic function $\gamma_\infty(\mu,\mathbf{t})$ is the pointwise limit of the functions $\gamma_\ell(\mu,\mathbf{t})$.

Proof. Let

$$\psi_{\ell,\mathbf{w}} := \bigwedge_{i=1}^{d}\bigwedge_{j=1}^{w_i}\phi_{\ell,i}(x_1,x_{n_{i,j}+1},\dots,x_{n_{i,j}+p_i-1})$$

(as in Lemma 27).

For each vector $\mathbf{w}\in\mathbb{N}^d$ the sequence $\left(\int_S k(\psi_{\ell,\mathbf{w}})\,\mathrm{d}\mu\right)_{\ell\in\mathbb{N}}$ is non-decreasing and bounded by 1 hence converging. It follows that the functions $\gamma_\ell(\mu,\mathbf{t})$ converge pointwise as $\ell\to\infty$ to

$$\gamma_\infty(\mu,\mathbf{t}) = \sum_{w_1\geq 0}\cdots\sum_{w_d\geq 0}\left(\lim_{\ell\to\infty}\int_S k(\psi_{\ell,\mathbf{w}})\,\mathrm{d}\mu\right)\prod_{j=1}^{d}\frac{(it_j)^{w_j}}{w_j!},$$

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

$$\mathbf{D}_\ell(\mu_{\mathbf{A}_n}) \xrightarrow{\mathcal{D}} \mathbf{D}_\ell(\mu_{\lim \mathbf{A}}) \qquad \text{as } n \to \infty$$
$$\mathbf{D}_\ell(\mu_{\lim \mathbf{A}}) \xrightarrow{\mathcal{D}} \mathbf{D}_\infty(\mu_{\lim \mathbf{A}}) \qquad \text{as } \ell \to \infty$$

However, although $D_\ell(\mu_{\mathbf{A}_n}) \xrightarrow{\mathcal{D}} \mathbf{D}_\infty(\mu_{\mathbf{A}_n})$ 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 σ-structures. The 1-*point random lift distribution* of $\mathbf{A}$ is the probability distribution over (isomorphism classes of) $\sigma^\bullet$-structures (where $\sigma^\bullet$ is the signature obtained from σ by adding a unary symbol M_1), corresponding to the marking a random elements X_1 of $\mathbf{A}$, drawn from A according to probability distribution $\nu_{\mathbf{A}}$. We denote by Π the map from the space $\mathrm{Rel}(\sigma)$ of isomorphism classes of finite σ-structures (with domain endowed with a probability measure) to the space $P(\mathrm{Rel}(\sigma))$ of probability distributions over $\mathrm{Rel}(\sigma)$, which maps a σ-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 σ-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 $f : A \backslash N_A \to B \backslash N_B$ preserving the measure (i.e. such that $\nu_{\mathbf{A}} = \nu_{\mathbf{B}} \circ f$) and all the relations both ways (i.e. $\mathbf{A} \models R(v_1, \ldots, v_n) \iff \mathbf{B} \models R(f(v_1), \ldots, f(v_n))$).

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

$$\iota^{\sigma^\bullet}_*(\zeta) = \zeta \circ (\iota^{\sigma^\bullet})^{-1}$$

is a continuous injection from $\mathrm{P}(\mathrm{Rel}(\sigma))$ to $\mathrm{P}(\mathfrak{M}_\sigma)$.

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* $\widetilde{\Pi} : \mathfrak{M}_\sigma \to \mathrm{P}(\mathfrak{M}_{\sigma^\bullet})$ *such that the following diagram commutes:*

$$\begin{array}{ccc} \mathrm{Rel}(\sigma) & \xrightarrow{\Pi} & \mathrm{P}(\mathrm{Rel}(\sigma^\bullet)) \\ {\scriptstyle \iota^\sigma}\uparrow & & \uparrow{\scriptstyle \iota^{\sigma^\bullet}_*} \\ \mathfrak{M}_\sigma & \xrightarrow[\widetilde{\Pi}]{} & \mathrm{P}(\mathfrak{M}_{\sigma^\bullet}) \end{array}$$

Proof. Consider an enumeration $\phi_1, \ldots, \phi_d, \ldots$ of local formulas with respect to signature $\sigma^\bullet$. To each formula ϕ_i with $p \geq 0$ free variables we associate the local formula ψ_i (with respect to signature σ) 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 $\sigma^\bullet$-structures $\mathbf{A}^+$ obtained by marking a single element $v \in A$ in a σ-structure $\mathbf{A}$. Then it holds

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

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

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

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

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

$$\mathrm{d}(\mu_1, \mu_2) \leq \alpha \quad \Longrightarrow \quad |F(\mu_1) - F(\mu_2)| \leq \epsilon.$$

Let $d = \lceil 1/\alpha \rceil$. Consider the continuous map $p : \mathfrak{M}_{\sigma^\bullet} \to [0,1]^d$ defined by

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

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 $(\langle \psi_1, \mathbf{A}_n \rangle_\bullet, \ldots, \langle \psi_d, \mathbf{A}_n \rangle_\bullet)$ converges in distribution. Thus for every box C_j the value

$$\begin{aligned} \int_{C_j} \mathrm{d}\zeta_n(\mu) &= \Pr[\mu \in C_j] && (\mu \text{ dist. wrt } \zeta_n) \\ &= \Pr[(\langle \phi_1, \mathbf{A}_n^+ \rangle, \ldots, \langle \phi_d, \mathbf{A}_n^+ \rangle) \in B_j] && (\mathbf{A}_n^+ \text{ dist. wrt } \Pi(\mathbf{A}_n)) \\ &= \Pr[(\langle \psi_1, \mathbf{A}_n \rangle_v, \ldots, \langle \psi_d, \mathbf{A}_n \rangle_v) \in B_j] && (v \text{ dist. wrt } \nu_{\mathbf{A}_n}) \end{aligned}$$

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 $\mathrm{d}(\mu_1, \mu_2) \leq 1/d$ hence for every $\mu \in C_j$ it holds $|F(\mu_1) - t_j| \leq \epsilon$. Thus it holds

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

As the sets C_j form a partition of $S_{\sigma^\bullet}$ and as ζ_n is a probability measure, we get

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

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

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

$$\widehat{\Pi}(\zeta) = \int_{\mathrm{Rel}(\sigma)} \Pi(\mathbf{A}) \, \mathrm{d}\zeta(\mathbf{A}),$$

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

$$\begin{array}{ccccc}
\mathbf{Rel}(\boldsymbol{\sigma}) & \xhookrightarrow{\iota} & \mathbf{P}(\mathbf{Rel}(\boldsymbol{\sigma})) & \xrightarrow{\widehat{\Pi}} & \mathbf{P}(\mathbf{Rel}(\boldsymbol{\sigma}^\bullet)) \\
\big\downarrow{\scriptstyle \iota} & & \big\downarrow{\scriptstyle \iota} & & \big\downarrow{\scriptstyle \iota} \\
\mathfrak{M}_\sigma & \xhookrightarrow{\iota} & \mathbf{P}(\mathfrak{M}_\sigma) & \xrightarrow{\widetilde{\widehat{\Pi}}} & \mathbf{P}(\mathfrak{M}_{\sigma^\bullet})
\end{array}$$

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. Although we do not have, in general, a nice limit structure for a local-convergent sequence of structures, we shall see that nevertheless 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 ϕ_d be the formula $\mathrm{dist}(x_1, x_2) \leq d$. Let $\mathbf{A}$ be a local convergent sequence of σ-structures, and let $D_{d,n} : A_n \to [0,1]$ be the random variable

$$D_{d,n}(v) = \langle \phi_d, \mathbf{A}_n \rangle_v = \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(v)).$$

As obviously ϕ_d implies ϕ_{d+1} it follows from Lemma 28 that there exists random variables D_d and D such that $D_{d,n} \xrightarrow{\mathcal{D}} D_d$ and $D_d \xrightarrow{\mathcal{D}} D$ (which are limits in distribution, for $n \to \infty$ and $d \to \infty$, 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 *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 $D_d \xrightarrow{\mathcal{D}} D$ (as $d \to \infty$) the functions F_d converge pointwise to F at every continuity point of F. Similarly, for each $d \in \mathbb{N}$, as $D_{n,d} \xrightarrow{\mathcal{D}} D_d$ (as $n \to \infty$) the functions $F_{n,d}$ converge pointwise to F_d at every continuity point of F_d. We define Λ_d (resp. Λ) as the (at most countable) set of discontinuities of F_d (resp. F), and let $\mathcal{R} = [0,1] \setminus \big(\Lambda \cup \bigcup_{d \in \mathbb{N}} \Lambda_d\big)$. In other words, $\mathcal{R}$ is the (cocountable) 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 $|F(t_j) - F_{d_1}(t_j)| < \epsilon$ and $|F_{d_i}(t_j) - F_{d_i,n}(t_j)| < \epsilon$ hold for $i, j \in \{1, 2\}$ then

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

Proof. The right inequality is obvious as $\Pr(t_1 < D_{d_1,n} \leq D_{d_2,n} \leq t_2) \leq \Pr(t_1 < D_{d_1,n} \leq t_2)$. For the left inequality, note that

$$\begin{aligned}\Pr(t_1 < D_{d_1,n} \leq D_{d_2,n} \leq t_2) &= \Pr(t_1 < D_{d_1,n} \leq t_2) - \Pr(t_1 < D_{d_1,n} \leq t_2 < D_{d_2,n}) \\ &\leq \Pr(t_1 < D_{d_1,n} \leq t_2) - (\Pr(D_{d_1,n} \leq t_2) - \Pr(D_{d_1,n} \leq t_2)) \\ &= (F_{d_1,n}(t_2) - F_{d_1,n}(t_1)) - (F_{d_1,n}(t_2) - F_{d_2,n}(t_2))\end{aligned}$$

□

We are now approaching the final steps of our cluster analysis. This is admittedly technical and we shall need further several lemmas in order to prove Theorem 1.

However the intuition for our proof is easy and can be outlined as follows: if we 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 λ hence an integral multiple of λ. 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 $p : [0, 1] \to [0, 1]$ defined by*

$$p(x) = F(x) - \lim_{\epsilon \to 0} F(x - \epsilon).$$

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 $D_d \xrightarrow{\mathcal{D}} D$ (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 $D_{n,d} \xrightarrow{\mathcal{D}} D_d$ (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 $v \in A_n$ such that $t_1 < D_{d,n}(v) \leq D_{2d,n}(v) \leq t_2$. For every $x \in \mathrm{N}^d_{\mathbf{A}_n}(v)$ it holds $\mathrm{N}^d_{\mathbf{A}_n}(x) \subseteq \mathrm{N}^{2d}_{\mathbf{A}_n}(v)$ hence $D_{d,n}(x) \leq D_{2d,n}(v) \leq t_2$. Also, $\mathrm{N}^{2d}_{\mathbf{A}_n}(x) \supseteq \mathrm{N}^d_{\mathbf{A}_n}(v)$ thus $D_{2d,n}(x) \geq D_{d,n}(v) > t_1$. As this holds for every $x \in \mathrm{N}^d_{\mathbf{A}_n}(v)$, we get $\mathrm{N}^d_{\mathbf{A}_n}(v) \subseteq \{x : t_1 < D_{2d,n}(x) \text{ and } D_{d,n}(x) \leq t_2\}$. Thus we have

$$\begin{aligned}
\nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(v)) - \Pr(t_1 < D_{2d,n} \leq t_2) &\leq \Pr(t_1 < D_{2d,n} \text{ and } D_{d,n} \leq t_2) - \Pr(t_1 < D_{2d,n} \leq t_2) \\
&\leq \Pr(D_{d,n} \leq t_2) - \Pr(D_{2d,n} \leq t_2) \\
&= F_{d,n}(t_2) - F_{2d,n}(t_2) \\
&< 4\epsilon.
\end{aligned}$$

Hence $\Pr(t_1 < D_{2d,n} \leq t_2) > \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(v)) - 4\epsilon > t_1 - 4\epsilon$. 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. □

Recall that Λ 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 λ. Indeed, we will prove later that $p(\lambda)/\lambda$ is the number of disjoint globular clusters with limit measure λ.

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 λ 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 $|F(t_1) - F_{kd}t_i| < \epsilon$ 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

$$\begin{aligned}
t_1 < D_{d,n}(v_1) \leq D_{4d,n}(v_1) \leq t_2, \\
t_1 < D_{d,n}(v_2) \leq D_{4d,n}(v_2) \leq t_2, \\
2d < \mathrm{dist}(v_1, v_2) \leq 3d.
\end{aligned}$$

Assume the contrary. Then $\mathrm{N}^{4d}_{\mathbf{A}_n}(v_1)$ contains the disjoint union of $\mathrm{N}^d_{\mathbf{A}_n}(v_1)$ and $\mathrm{N}^d_{\mathbf{A}_n}(v_2)$ thus

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{4d}_{\mathbf{A}_n}(v_1)) > 2t_1 > t_1 + \lambda - \epsilon = (t_1 + \epsilon) + (\lambda - 2\epsilon) > t_2,$$

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 $\mathrm{N}^d_{\mathbf{A}_n}(v)$ and $\mathrm{N}^d_{\mathbf{A}_n}(v')$ do not intersect hence

$$1 \geq \nu_{\mathbf{A}_n}\big(\bigcup_{v \in S} \mathrm{N}^d_{\mathbf{A}_n}(v)\big) = \sum_{v \in S} D_{d,n}(v) > t_1|S| > (\lambda - \epsilon)|S|.$$

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 $\bigcup_{v\in S} \mathrm{N}^{2d}_{\mathbf{A}_n}(v) = \mathrm{N}^{2d}_{\mathbf{A}_n}(S)$. It follows that $\Pr(t_1 < D_{d,n} \leq D_{4d,n} \leq t_2) \leq t_2|S|$. Also,

$$\Pr(t_1 < D_{2d,n} \leq t_2) \geq \nu_{\mathbf{A}_n}(\mathrm{N}^d_{\mathbf{A}_n}(S)) = \sum_{s\in S} D_{d,n}(s) > t_1|S| > \lambda|S| - \epsilon/(\lambda-\epsilon).$$

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 ϵ_z. For $z \in \mathbb{N}$, we define

$$\epsilon_z = 2^{-z}. \tag{6}$$

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

$$z_0(\lambda) = \lceil 5 - 2\log_2 \lambda \rceil. \tag{7}$$

(Thus $\epsilon_{z_0(\lambda)} \leq \lambda^2/32$.)

Definition of $\alpha_z(\lambda)$ and $\beta_z(\lambda)$. We define

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

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

$$|\beta_z(\lambda) - \alpha_z(\lambda)| < \epsilon_z. \tag{8}$$

Definition of $\delta_z(\lambda)$. As $D_d \xrightarrow{\mathcal{D}} D$ (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

$$|F_d(\alpha_z(\lambda)) - F(\alpha_z(\lambda))| < \epsilon_z \tag{9}$$

$$|F_d(\beta_z(\lambda)) - F(\beta_z(\lambda))| < \epsilon_z \tag{10}$$

Definition of $\eta_z(\lambda)$. As $D_{n,d} \xrightarrow{\mathcal{D}} D_d$ (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

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

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

We now define some sequences of sets. The sets $Z^{\lambda,z}_n$ will anticipate our construction of globular clusters, by giving a rough approximate of them. Then the set S^λ_n will collect a "center" for each of the "component" of size λ.

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

- If $n < \eta_z$ then $Z^{\lambda,z}_n = \emptyset$;

- Otherwise, $Z_n^{\lambda,z}$ is the set of all elements of A_n such that

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

$$D_{\delta_{z'},n}(v) > \alpha_{z'}(\lambda) \qquad (\forall z' \in \{z_0(\lambda), \dots, z\}) \tag{14}$$

Definition of S_n^λ. We define S_n^λ as a maximal set of vertices $v \in Z_n^{\lambda,z}$, 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

$$Z_n^{\lambda,z_0(\lambda)} \supseteq Z_n^{\lambda,z_0(\lambda)+1} \supseteq \cdots \supseteq Z_n^{\lambda,z} \supsetneq Z_n^{\lambda,z+1} = Z_n^{\lambda,z+2} = \cdots = \emptyset. \tag{15}$$

Remark 9. According to the definitions of δ_z and η_z, it holds

$$|F_{8\delta_z(\lambda),n}(\beta_z(\lambda) - F(\beta_z(\lambda)| < 2\epsilon_z$$

Remark 10. If $z' < z''$ then, according to Lemma 29 it holds

$$|\Pr(\alpha_{z'}(\lambda) < D_{\delta_{z'}(\lambda),n} \leq D_{\delta_{z''}(\lambda),n} \leq \alpha_{z''}(\lambda))| < 4\epsilon_{z'}$$

Thus

$$\Pr(D_{\delta_z(\lambda),n} > \alpha_z(\lambda)) - \Pr\Big(\bigwedge_{z'=z_0}^{z} D_{\delta_{z'}(\lambda),n} > \alpha_{z'}(\lambda)\Big) < 4 \sum_{z'=z_0}^{z} \epsilon_{z'} = 2^{2-z_0}.$$

It follows that

$$\nu_{\mathbf{A}_n}(Z_n^{\lambda,z}) \geq p(\lambda) - 4\epsilon_z - 2^{2-z_0}.$$

We now prove that, as wanted, the number of elements of S_n^λ is (for sufficiently large λ) the anticipated number of globular clusters of size λ.

Lemma 32. *For every $\lambda \in \mathcal{R}$ and every $n \geq \eta_{\lceil \lambda^{-1} \rceil}$ it holds $|S_n^\lambda| = p(\lambda)/\lambda$.*

Proof. Note that obviously, as $\nu_{\mathbf{A}_n}(\mathrm{N}_{\mathbf{A}_n}^{7\delta_z(\lambda)}(s)) < \lambda + \epsilon_z$ holds for every $s \in S_n^\lambda$, we get

$$|S_n^\lambda| \geq \frac{|Z_n^{\lambda,z}|}{\lambda + \epsilon_z} \geq \frac{p(\lambda) - 2^{2-z} - 2^{2-z_0(\lambda)}}{\lambda + 2^{-z}} = \frac{p(\lambda)}{\lambda} - \frac{2^{-z}/\lambda - 2^{2-z} - 2^{2-z_0(\lambda)}}{\lambda - 2^{-z}} > \frac{p(\lambda)}{\lambda} - 1,$$

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 $D_{3\delta_z(\lambda),n}(s) \leq \beta_z(\lambda)$ thus for every $v \in \mathrm{N}_{\mathbf{A}_n}^{\delta_z(\lambda)}(S_n^\lambda)$ it holds $\alpha_z(\lambda) < D_{2\delta_z(\lambda),n}(s) \leq \beta_z(\lambda)$ thus

$$\begin{aligned}(\lambda - \epsilon_z)|S_n^\lambda| &< \nu_{\mathbf{A}_n}(\mathrm{N}_{\mathbf{A}_n}^{\delta_z(\lambda)}(S_n^\lambda)) \\ &\leq \Pr(\alpha_z(\lambda) < D_{2\delta_z(\lambda),n}(s) \leq \beta_z(\lambda)) \\ &< p(\lambda) + 4\epsilon_z\end{aligned}$$

hence

$$|S_n^\lambda| < \frac{p(\lambda)}{\lambda} + 1.$$

Altogether, it follows that $|S_n^\lambda| = p(\lambda)/\lambda$. □

We are now ready to define sets gathering all the "components" with limit measure λ. We will prove that they define universal clusters.

Definition of C_n^λ. For $\lambda \in \Lambda$ and $n \in \mathbb{N}$ we define

$$C_n^\lambda = \begin{cases} \emptyset, & \text{if } n < \eta_{z_0(\lambda)} \\ \mathrm{N}_{\mathbf{A}_n}^{2\delta_z}(S_n^\lambda), & \text{otherwise, if } z \text{ is such that } \eta_z \leq n < \eta_{z+1} \end{cases} \tag{16}$$

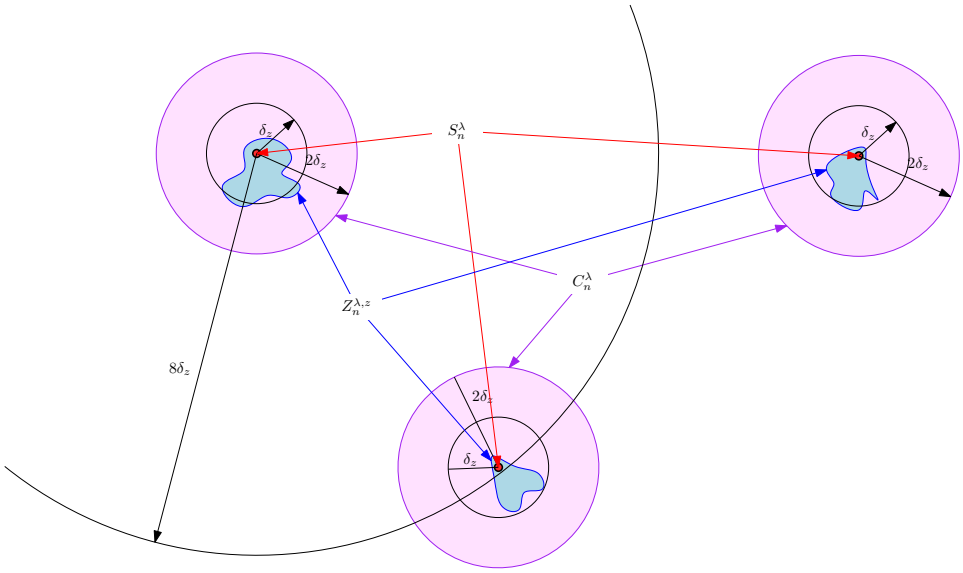

FIGURE 4. General aspects of the sets $Z_n^{\lambda,z}, S_n^\lambda$, and C_n^λ

The sets C_n^λ will be the building block for the construction of our clusters. Lemmas 33 to 38 will be used to prove that the sequences C^λ define a clustering of $\mathbf{A}$ into countably many universal clusters plus a residual cluster. The general aspects of the sets $Z_n^{\lambda,z}, S_n^\lambda$, and C_n^λ we tried to visualize by Fig. 4.

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 $v \in Z_n^\lambda \setminus C_n^\lambda$. By the maximality of S_n^λ, we get that v is at distance at most $7\delta_z$ from some element $u \in S_n^\lambda$. Moreover, $\mathrm{dist}(u,v) > 2\delta_z$ as $v \notin \mathrm{N}^{2\delta_z}_{\mathbf{A}_n}(S_n^\lambda)$. Then $\mathrm{N}^{8\delta_z}_{\mathbf{A}_n}(v)$ contains the disjoint union of $\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(v)$ and $\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(u)$ thus

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{8\delta_z}_{\mathbf{A}_n}(v)) > 2\alpha_z(\lambda) > \alpha_z(\lambda) + \lambda - \epsilon_2 = (\alpha_z(\lambda) + \epsilon_z) + (\lambda - 2\epsilon_z) > \beta_z(\lambda),$$

contradicting $D_{8\delta_z,n}(v) \leq \beta_z(\lambda)$. □

We now prove that our sets C_n^λ are pairwise disjoint.

Lemma 34. *Let $\lambda < \lambda'$ be two elements of Λ, let $z \geq \max(z_0(\lambda), \lceil 1 - \log_2(\lambda' - \lambda)\rceil)$, 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 $v \in \mathrm{N}^{2\delta_z}_{\mathbf{A}_n}(u) \cap \mathrm{N}^{2\delta_z}_{\mathbf{A}_n}(u')$ hence $\mathrm{dist}(u,u') \leq 4\delta_z$. It follows that $\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(u') \subseteq \mathrm{N}^{8\delta_z}_{\mathbf{A}_n}(u)$ 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^λ 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 $|\nu_{\mathbf{A}_n}(C_n^\lambda) - p(\lambda)| < 2^{-z} p(\lambda)/\lambda$.

Proof. Note that

$$\nu_{\mathbf{A}_n}(C_n^\lambda) = \sum_{s\in S_n^\lambda} \nu_{\mathbf{A}_n}(\mathrm{N}^{2\delta_z}_{\mathbf{A}_n}(s)).$$

Hence, as $|S_n^\lambda| = p(\lambda)/\lambda)$ it holds $|\nu_{\mathbf{A}_n}(C_n^\lambda) - p(\lambda)| < 2^{-z}p(\lambda)/\lambda$. □

The next lemma not only shows that the outer boundary of C^λ is negligible (what is required in order for C^λ 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 $n \geq \eta_{\lceil \lambda^{-1}\rceil}$, and let $z \in \mathbb{N}$ be such that $\eta_z \leq n < \eta_{z+1}$. Then it holds*

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} C_n^\lambda)) < 2^{1-z}p(\lambda)/\lambda.$$

In particular, $\partial_{\mathbf{A}}\mathsf{C}^\lambda \approx \mathsf{0}$.

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

$$\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} C_n^\lambda) = \biguplus_{v\in S_n^\lambda} \left(\mathrm{N}^{3\delta_z}_{\mathbf{A}_n}(v) \setminus \mathrm{N}^{\delta_z}_{\mathbf{A}_n}(v)\right)$$

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

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(v)) = D_{\delta_z,n}(v) > \alpha_z(\lambda)$$
$$\nu_{\mathbf{A}_n}(\mathrm{N}^{3\delta_z}_{\mathbf{A}_n}(v)) \leq D_{8\delta_z,n}(v) \leq \beta_z(\lambda)$$

Hence

$$\nu_{\mathbf{A}_n}\left(\mathrm{N}^{3\delta_z}_{\mathbf{A}_n}(v) \setminus \mathrm{N}^{\delta_z}_{\mathbf{A}_n}(v)\right) < \epsilon_z$$

Thus

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} C_n^\lambda)) < |S_n^\lambda|\epsilon_z < 2\epsilon_z p(\lambda)/\lambda$$

□

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*

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

Then

$$\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(W_n)) \leq 2^{-z}(1 + 3/\lambda).$$

In particular, $\mathsf{W} \supseteq \partial_{\mathbf{A}}\left(\bigcup_{\lambda\in\Lambda} \mathsf{C}^\lambda\right)$ and $\mathsf{W} \approx \mathsf{0}$.

Proof. Let $F_n = \{v : D_{\delta_z,n}(v) > \alpha_z(\lambda)\}$. Then $\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(F_n) \subseteq \{v : D_{\delta_2 z,n}(v) > \alpha_z(\lambda)\}$. Hence $\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(F_n)) \leq 1 - F_{2\delta_z,n}(\alpha_z(\lambda))$. It follows that

$$\begin{aligned}\nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(W_n)) &\leq \nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(F_n)) - \sum_{\lambda'\geq\lambda} \nu_{\mathbf{A}_n}(C_n^{\lambda'}) + \sum_{\lambda'\geq\lambda} \nu_{\mathbf{A}_n}(\mathrm{N}^{\delta_z}_{\mathbf{A}_n}(\partial_{\mathbf{A}_n} C_n^{\lambda'})) \\ &\leq \epsilon_z + \frac{2^{-z}}{\lambda} + \frac{2^{1-z}}{\lambda}.\end{aligned}$$

□

We now are ready for our last lemma needed to prove that the sequences C^λ define a clustering of $\mathbf{A}$ into countably many universal clusters plus a residual cluster.

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

Proof. Let ϕ be an r-local strongly local formula with free variables $x_1, \dots, x_q$. For $d \in \mathbb{N}$ let Ψ_d be the following formula with $q+1$ free variables

$$\Psi_d : \quad \phi(x_2, \dots, x_{q+1}) \wedge \bigwedge_{i=2}^{q+1} \operatorname{dist}(x_1, x_i) \le d.$$

Note that if $d_1 < d_2$ and $v \in A_n$ then

$$0 \le \langle \Psi_{d_2}, \mathbf{A}_n \rangle_v - \langle \Psi_{d_1}, \mathbf{A}_n \rangle_v \le q\big(D_{d_2,n}(v) - D_{d_1,n}(v)\big).$$

For $\lambda \in \Lambda$ we consider an integer z_1 such that $\delta_{z_1} > r$ and $z_1 \ge z_0(\lambda)$, an integer $z \ge z_1$ and $\eta_z \le n < \eta_{z+1}$. Then the following holds: for every $s \in S_n^\lambda$ and every $x \in \mathrm{N}_{\mathbf{A}_n}^{\delta_{z_1}}(s)$, it holds $\mathrm{N}_{\mathbf{A}_n}^d(x) \subseteq \mathrm{N}_{\mathbf{A}_n}^{d+\delta_{z_1}}(s)$ and $\mathrm{N}_{\mathbf{A}_n}^d(s) \subseteq \mathrm{N}_{\mathbf{A}_n}^{d+\delta_{z_1}}(x)$ we get

$$\langle \Psi_{d-\delta_{z_1}}, \mathbf{A}_n \rangle_s \le \langle \Psi_d, \mathbf{A}_n \rangle_x \quad \text{and} \quad \langle \Psi_d, \mathbf{A}_n \rangle_x \le \langle \Psi_{d+\delta_{z_1}}, \mathbf{A}_n \rangle_s.$$

It follows that

$$\begin{aligned}
\langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_x - \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s &\le \langle \Psi_{3\delta_{z_1}}, \mathbf{A}_n \rangle_s - \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s \\
&\le q \Pr(2\delta_{z_1} \le \operatorname{dist}(x,s) \le 3\delta_{z_1}) \\
&\le q \left(D_{8\delta_z,n}(s) - D_{\delta_{z_1},n}(s)\right) \\
&< q(\beta_z(\lambda) - \alpha_{z_1}(\lambda)) \\
&< q(\epsilon_{z_1} + \epsilon_z)
\end{aligned}$$

and

$$\begin{aligned}
\langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s - \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_x &\le \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s - \langle \Psi_{\delta_{z_1}}, \mathbf{A}_n \rangle_s \\
&\le q \Pr(\delta_{z_1} \le \operatorname{dist}(x,s) \le 2\delta_{z_1}) \\
&< q(\epsilon_{z_1} + \epsilon_z).
\end{aligned}$$

Thus

$$|\langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_x - \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s| < q(\epsilon_{z_1} + \epsilon_z).$$

Also,

$$\begin{aligned}
|\langle \Psi_{2\delta_z}, \mathbf{A}_n \rangle_s - \langle \Psi_{2\delta_{z_1}}, \mathbf{A}_n \rangle_s| &\le q(D_{\delta_z,n}(s) - D_{\delta_{z_1},n}(s)) \\
&\le q(D_{8\delta_z,n}(s) - D_{\delta_{z_1},n}(s)) \\
&\le q(\beta_z(\lambda) - \alpha_{z_1}(\lambda)) \\
&< q(\epsilon_{z_1} + \epsilon_z).
\end{aligned}$$

Moreover,

$$|\langle \Psi_{2\delta_z}, \mathbf{A}_n \rangle_s - \langle \Psi_{2\delta_z}, \mathbf{A}_n - \partial_{\mathbf{A}_n} C_n^\lambda \rangle_s| < 4q\epsilon_z p(\lambda)/\lambda.$$

and

$$\langle \phi, \mathbf{A}_n[C_n^\lambda] \rangle = \frac{\sum_{s \in S_n^\lambda} \langle \Psi_{2\delta_z}, \mathbf{A}_n - \partial_{\mathbf{A}_n} C_n^\lambda \rangle_s}{\nu_{\mathbf{A}_n}(C_n^\lambda)^p}.$$

Thus, as

$$|\nu_{\mathbf{A}_n}(C_n^\lambda) - \lambda| < \epsilon_z p(\lambda)/\lambda,$$

it holds

$$\begin{aligned}\mathbb{E}[\langle\Psi_{2\delta_{z_1}}, A_n\rangle_v \mathbf{1}_{Z_n^{\lambda,z_1}}(v)] &= \frac{\sum_{v\in Z_n^{\lambda,z_1}} \nu_{\mathbf{A}_n}(v)\,\langle\Psi_{2\delta_{z_1}}, \mathbf{A}_n\rangle_v}{\nu_{\mathbf{A}_n}(Z_n^{\lambda,z_1})}\\ &\approx \frac{1}{\lambda}\sum_{v\in C_n^\lambda}\langle\Psi_{2\delta_z}, \mathbf{A}_n\rangle_v\\ &\approx \sum_{s\in S_n^\lambda}\langle\Psi_{2\delta_z}, \mathbf{A}_n\rangle_s\\ &\approx \lambda^p\langle\phi, \mathbf{A}_n[C_n^\lambda]\rangle\end{aligned}$$

Let $H_{z_1,n}$ be the (multivariate) cumulative distribution function of

$$(\langle\Psi_{\delta_{z_1}}, \mathbf{A}_n\rangle_\bullet, 1-D_{\delta_{z_0(\lambda)}}, \dots, 1-D_{\delta_{z_1}}, D_{8\delta_{z_1}}).$$

According to its definition we have $v\in Z_n^{\lambda,z_1}$ if and only if

$(1-D_{\delta_{z_0(\lambda)}}, \dots, 1-D_{\delta_{z_1}}, D_{8\delta_{z_1}}) \in [0,1-\alpha_{z_0(\lambda)}(\lambda)]\times\cdots\times[0,1-\alpha_{z_1}(\lambda)]\times[0,\beta_{z_1}(\lambda)]$.

Thus

$\Pr[\langle\Psi_{\delta_{z_1}}, \mathbf{A}_n\rangle_v \le x \text{ and } v\in Z_n^{\lambda,z_1}] = H_{n,z_1}(x, 1-\alpha_{z_0(\lambda)}(\lambda), \dots, 1-\alpha_{z_1}(\lambda), \beta_{z_1}(\lambda))$.

It follows that

$$\begin{aligned}\mathbb{E}[\langle\Psi_{2\delta_{z_1}}, A_n\rangle_v \mathbf{1}_{Z_n^{\lambda,z_1}}(v)] &= \int_0^1 \Pr[\langle\Psi_{\delta_{z_1}}, \mathbf{A}_n\rangle_v \le x \text{ and } v\in Z_n^{\lambda,z_1}]\,\mathrm{d}x\\ &= \int_0^1 1-H_{n,z_1}(x, 1-\alpha_{z_0(\lambda)}(\lambda), \dots, 1-\alpha_{z_1}(\lambda), \beta_{z_1}(\lambda))\,\mathrm{d}x.\end{aligned}$$

According to Lemma 27 there exists a (vector) random variable $\mathbf{V}_{z_1}$ such that

$$(\langle\Psi_{\delta_{z_1}}, \mathbf{A}_n\rangle_\bullet, 1-D_{\delta_{z_0(\lambda)}}, \dots, 1-D_{\delta_{z_1}}, D_{8\delta_{z_1}}) \xrightarrow{\mathcal{D}} \mathbf{V}_{z_1}.$$

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

$$\lim_{n\to\infty}\mathbb{E}[\langle\Psi_{2\delta_{z_1}}, A_n\rangle_v \mathbf{1}_{Z_n^{\lambda,z_1}}(v)] = \int_0^1 1-H(x, 1-\alpha_{z_0(\lambda)}(\lambda), \dots, 1-\alpha_{z_1}(\lambda), \beta_{z_1}(\lambda))\,\mathrm{d}x.$$

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

We are now ready to prove our first clustering result:

Lemma 39. *Let $\mathbf{A}$ be a local convergent sequence of σ-structures. Let σ^+ be the signature obtained from σ 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 $\lambda_1>\lambda_2>\dots$ are the elements of Λ order in decreasing order) and by M_0 the sequence of sets*

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

we obtain clustering $\mathrm{L}(\mathbf{A})$ of $\mathbf{A}$ with the following properties:

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

Proof. That $\mathrm{L}(A)$ is clustering follows from Lemma 22. That $\mathbf{C}^\lambda$ 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 $\mathbf{A}$. 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 σ-structures. Then there exists a signature σ^+ (obtained from σ 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 $\lambda_1 > \lambda_2 > \dots$ a positive reals and a clustering $\mathrm{L}(\mathbf{A})$ of $\mathbf{A}$ with the following properties:*

- *For every $i \in \mathbb{N}$, $1 \leq j \leq a_i$, and $1 \leq k \leq b_{i,j}$, $\mathsf{G}^{\mathsf{i,j,k}} = M_{i,j,k}(\mathrm{L}(\mathbf{A}))$ is a globular cluster of $\mathbf{A}$ such that $\lim \nu_{\mathbf{A}}(\mathsf{G}^{i,j,k}) = \lambda_i$, that is a cluster such that for every positive real ϵ there is an integer d which satisfies*
$$\lambda_i - \epsilon < \liminf_{n\to\infty} \max_{v_n \in G_n^{i,j,k}} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{v_n}) \leq \lim_{n\to\infty} \nu_{\mathbf{A}_n}(G_n^{i,j,k}) = \lambda_i.$$
- *$\mathsf{R} = M_R(\mathrm{L}(\mathbf{A}))$ is a residual cluster of $\mathbf{A}$, that is a cluster such that for every integer d it holds*
$$\limsup_{n\to\infty} \max_{v_n \in G_n^{i,j,k}} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{v_n}) = 0.$$
- *The sequence S is negligible, that is such that for every integer d it holds*
$$\limsup_{n\to\infty} \nu_{\mathbf{A}_n}(\mathrm{N}^d_{S_n}) = 0.$$
- *The marks partition the sets A_n is a stable way, that is*
$$\lim \nu_{\mathbf{A}}(\mathsf{R}) + \sum_{i\geq 1} \lim \nu_{\mathbf{A}}(\mathsf{G}^{i,j,k}) = 1.$$
- *Clusters $\mathsf{G}^{i,j,k}$ and $\mathsf{G}^{i',j',k'}$ are interweaving (i.e. $\mathsf{G}^{i,j,k} \between \mathsf{G}^{i',j',k'}$) if and only if $i = i'$ and $j = j'$.*
- *The clusters $\bigcup_{k=1}^{b_{i,j}} \mathsf{G}^{i,j,k}$ (grouping interweaving clusters) are universal.*
- *The number $N_i = \sum_{j=1}^{a_i} b_{i,j}$ of clusters with limit measure λ_i is*
$$N_i = \frac{1}{\lambda_i} \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} \left[\sum_{w\geq 1} \left(\lim_{d\to\infty} \int_{S_\sigma} k(\psi_{d,w})\, \mathrm{d}\mu \right) \frac{(is)^w}{w!} \right] e^{-i\lambda_i s}\, \mathrm{d}s,$$
where $\psi_{d,w}$ is the formula
$$\psi_{d,w}(x_1, \dots, x_{w+1}) := \bigwedge_{i=1}^{w} \mathrm{dist}(x_1, x_i) \leq d.$$

Proof. By construction, the number of connected components of $\mathbf{A}_n[C_n^\lambda]$ is asymptotically $p(\lambda)/\lambda$ and each of these connected components has asymptotically measure λ. Let $\mathbf{B}_{n,1}, \dots, \mathbf{B}_{n,k_n}$ be the connected components of $\mathbf{A}_n[C_n^\lambda]$. If there is a local formula ϕ such that

$$\lim_{n\to\infty} \min_i \langle \phi, \mathbf{B}_i \rangle \neq \lim_{n\to\infty} \min_i \langle \phi, \mathbf{B}_i \rangle$$

we can break C^λ into smaller universal clusters. At the end of the day, we get a clustering of $\mathbf{A}$ 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 it the disjoint union of k_i interweaving clusters.

The statement giving the number N_i of clusters with measure λ_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 $\gamma_\infty(\mu, t)$ associated to the formulas $\mathrm{dist}(x_1, x_2) \leq d$ by Lemma 2. □

A direct consequence of Theorem 7 stands in the following complete characterization of the globular clusters of a local convergent sequence.

Theorem 8. *We have the following complete characterization of the globular clusters of a local convergent sequence* $\mathbf{A}$*: For a sequence* X *of subsets of* $\mathbf{A}$ *the following are equivalent*

(1) X *is a globular cluster of* $\mathbf{A}$*;*
(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*

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

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

$$\sum_{i>i_0} \sum_{j=1}^{a_i} \nu_{\mathbf{A}_n}(Z_n^{i,j}) < \epsilon.$$

Then notice that $\mathsf{X} \cap \mathsf{R} \approx \mathsf{0}$ as R is residual and X is not. According to Lemma 21, for each i, j, k it either holds $\mathsf{X} \cap \mathsf{G}^{i,j,k} \approx \mathsf{0}$ or $\mathsf{X} \between \mathsf{G}^{i,j,k}$. 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 $\mathsf{X} \cap \mathsf{Z}^{i,j} \approx 0$ it holds

$$\nu_{\mathbf{A}_n}(R_n \cap X_n) < \epsilon \text{ and } \nu_{\mathbf{A}_n}(R_n \cap Z_n^{i,j}) < \frac{\epsilon}{\sum_{i=1}^{i_0} a_i}.$$

Then, letting $\epsilon < \lim \nu_{\mathbf{A}}(\mathsf{X})/4$ we get that there exists integers i, j, k such that $\mathsf{X} \between \mathsf{G}^{i,j,k}$. It follows that $\mathsf{X} \between \mathsf{G}^{i',j',k'}$ if and only if $i = i'$ and $j = j'$. Thus for every $(i', j') \neq (i, j)$ it holds $\mathsf{X} \cap \mathsf{Z}^{i,j} \approx 0$, and thus it holds $\liminf \nu_{\mathbf{A}}(\mathsf{X} \cap \mathsf{Z}^{i,j}) \geq 1 - 3\epsilon$. Letting $\epsilon \to 0$ we get that $\mathsf{Z}^{i,j} \setminus \mathsf{X}$ is negligible. As X is globular and as $\mathsf{Z}^{i,j}$ 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 $\mathsf{Y} \approx \mathsf{X}$. □

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 [10]. This will be the subject of a forthcoming paper.

References

[1] Aldous, D.: Representations for partially exchangeable arrays of random variables. J. Multivar. Anal. **11**, 581–598 (1981)

[2] Alon, N.: Eigenvalues and expanders. Combinatorica **6**(2), 83–96 (1986)

[3] Benjamini, I., Schramm, O.: Recurrence of distributional limits of finite planar graphs. Electron. J. Probab. **6**(23), 13pp (2001)

[4] Bobkov, S., Houdré, C., Tetali, P.: λ_∞, vertex isoperimetry and concentration. Combinatorica **20**(2), 153–172 (2000)

[5] Bollobás, B., Riordan, O.: Sparse graphs: metrics and random models. Random Structures & Algorithms **39**(1), 1–38 (2011)

[6] Borgs, C., Chayes, J., Lovász, L., Sós, V., Vesztergombi, K.: Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing. Adv. Math. **219**(6), 1801–1851 (2008). DOI 10.1016/j.aim.2008.07.008

[7] Elek, G., Szegedy, B.: Limits of hypergraphs, removal and regularity lemmas. A non-standard approach. arXiv:0705.2179v1 [math.CO] (2007)

[8] Everitt, B., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley (2011)

[9] Gaifman, H.: On local and non-local properties. In: Proceedings of the Herbrand Symposium, Logic Colloquium '81 (1982)

[10] Hatami, H., Lovász, L., Szegedy, B.: Limits of locally–globally convergent graph sequences. Geometric and Functional Analysis **24**(1), 269–296 (2014)

[11] Hoover, D.: Relations on probability spaces and arrays of random variables. Tech. rep., Institute for Advanced Study, Princeton, NJ (1979)

[12] Lovász, L., Szegedy, B.: Limits of dense graph sequences. J. Combin. Theory Ser. B **96**, 933–957 (2006)

[13] Nešetřil, J., Ossona de Mendez, P.: A model theory approach to structural limits. Commentationes Mathematicæ Universitatis Carolinæ **53**(4), 581–603 (2012)

[14] Nešetřil, J., Ossona de Mendez, P.: Modeling limits in hereditary classes: Reduction and application to trees (2015). Submitted

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