Abstract

When studying flocking/swarming behaviors in animals one is interested in quantifying and comparing the dynamics of the clustering induced by the coalescence and disbanding of animals in different groups.

Motivated by this, we propose a summarization of time-dependent metric data which captures their time-dependent clustering features which we call formigrams. These set-valued functions generalize the notion of dendrogram, a prevalent object in the context of hierarchical clustering.

Also, we define a metric on formigrams for quantifying the degree of structural difference between any two given formigrams. In particular, the restriction of this metric to the collection of dendrograms recovers twice the Gromov-Hausdorff distance between the ultrametric spaces associated to the dendrograms. This fact enables us to show that constant factor approximations to the metric on formigrams cannot be obtained in polynomial time.

Finally, we investigate a sufficient condition for time-dependent metric spaces to be summarized into formigrams. In addition, we prove that this summarization process is stable under perturbations in the input time-dependent metric data.

1 Introduction

Given data represented as a static finite metric space \((X, d_X)\), a hierarchical clustering method finds a hierarchical family of partitions that captures multi-scale features present in the dataset. These hierarchical families of partitions are called dendrograms and their visualization is straightforward (see figure on the left).

We now turn our attention to a problem of characterizing dynamic data. We model dynamic datasets as time varying finite metric spaces and study a simple generalization of the notion of dendrogram which we call formigrarn - a combination of the words formicarium and diagram (see figure on the right).

Whereas dendrograms are useful for modeling situations when data points aggregate along a certain scale parameter, formigrams are better suited for representing phenomena when data points may also separate or disband and then regroup at different parameter values. One motivation for considering this scenario comes from the study and characterization of flocking/swarming/herding behavior of animals [1, 10, 11, 12, 19, 21, 24, 28], convoys [14], moving clusters [15], or mobile groups [13, 29].

Related work. Let \(X\) be a set of points having piecewise linear trajectories with time-stamped vertices in Euclidean space \(\mathbb{R}^d\). Buchin and et al. [3] provided explicit algorithms for studying the grouping structure of \(X\). This was subsequently enriched in [18, 20, 26, 27]. From the set \(X\), the authors of [3] construct a Reeb graph-like structure \(R_X\) which is closely related to the formigram derived from \(X\) that we introduce (Section 3 and Theorem 4). The edges of \(R_X\) are labeled by maximal groups, and they call \(R_X\) together with these labels the trajectory grouping structure of \(X\), enabling the visualization of the life span of maximal groups.

Our contributions.

1. We generalize dendrograms to formigrams for the analysis of clustering features of dynamic data, such as dynamic metric spaces or dynamic graphs.

2. Any dendrogram over a finite set \(X\) induces an ultrametric on \(X\) [7]. Therefore, one can quantify the structural difference between any two dendrograms by computing the Gromov-Hausdorff distance between their two induced ultrametrics [7]. We propose a distance \(d^F\) between formigrams which generalizes the method above for comparing two dendrograms (Theorems 1 and 2). The desire to obtain such a precise quantification of the difference between two dynamic clusterings was already made explicit in [3, Section 6]. Also, we show that constant factor approximations to \(d^F\) cannot be obtained in polynomial time (Theorem 3).

\footnote{A formicarium or ant farm is an enclosure for keeping ants under semi-natural conditions [30].}

---

Formigrams: Clustering Summaries of Dynamic Data*

Woojin Kim†
Facundo Mémoli‡

CCCG 2018, Winnipeg, Canada, August 8–10, 2018

*This work was partially supported by NSF grants IIS-1422400, CCF-1526513, and DMS-1723003.
†Department of Mathematics, The Ohio State University, kim.5235@osu.edu
‡Department of Mathematics, The Ohio State University, memoli@math.osu.edu
3. As an application, we propose a method for turning any (tame) dynamic metric data into a formigram. This method is closely related to the construction of trajectory grouping structures \cite{3}. In particular, this method turns out to be stable under perturbations in the input dynamic metric data under a certain notion of distance between DMSs that we introduce (Theorem 5).

2 Background

2.1 Dendrograms and treegrams

Partitions and sub-partitions. Let \( X \) be a non-empty finite set. We will call any partition \( P \) of a subset \( X' \) of \( X \) a sub-partition of \( X \) (in particular, any partition of the empty set is defined as the empty set). In this case we call \( X' \) the underlying set of \( P \).

1. By \( \mathcal{P}_{\text{sub}}(X) \), we denote the set of all sub-partitions of \( X \), i.e.
\[
\mathcal{P}_{\text{sub}}(X) := \{ P : \exists X' \subset X \text{, } P \text{ is a partition of } X' \}.
\]

2. By \( \mathcal{P}(X) \), we denote the subcollection of \( \mathcal{P}_{\text{sub}}(X) \) consisting solely of partitions of the whole \( X \).

Given \( P, Q \in \mathcal{P}_{\text{sub}}(X) \), by \( P \leq Q \) we mean “\( P \) is finer than or equal to \( Q \)”, i.e. for all \( B \in P \), there exists \( C \in Q \) such that \( B \subset C \). For example, let \( X = \{x_1, x_2, x_3\} \) and consider the sub-partitions \( P := \{\{x_1, x_2\}\} \) and \( Q := \{\{x_1, x_2\}, \{x_3\}\} \) of \( X \). Then, it is easy to see that in this case \( P \leq Q \).

Dendrograms. A dendrogram over a finite set \( X \) is any function \( \theta_X : R_+ \to \mathcal{P}(X) \) such that the following properties hold: (1) \( \theta_X(0) = \{\{x\} : x \in X\} \), (2) if \( t_1 \leq t_2 \), then \( \theta_X(t_1) \subseteq \theta_X(t_2) \), (3) there exists \( T > 0 \) such that and \( \theta_X(t) = \{X\} \) for \( t \geq T \), (4) for all \( t \) there exists \( \epsilon > 0 \) s.t. \( \theta_X(s) = \theta_X(t) \) for \( s \in [t, t + \epsilon] \) (right-continuity). See Figure 1 for an example.

\[\]

Figure 1: A dendrogram \( \theta_X \) over the set \( X = \{x_1, x_2, x_3, x_4\} \). Notice that \( \theta_X(0) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}\} \) and \( \theta_X(t) = \{X\} \) for all \( t \in [T, \infty) \).

Treegrams. Dendrograms can be generalized to treegrams, a visual representation for hierarchical clustering of networks \cite{24}. A treegram over a finite set \( X \) is any function \( \theta_X : R \to \mathcal{P}_{\text{sub}}(X) \) such that the following properties hold: (1) if \( t_1 \leq t_2 \), then \( \theta_X(t_1) \subseteq \theta_X(t_2) \), (2) (boundedness) there exists \( T > 0 \) such that \( \theta_X(t) = \{X\} \) for \( t \geq T \) and \( \theta_X(t) \) is empty for \( t \leq -T \), (3) for all \( t \) there exists \( \epsilon > 0 \) s.t. \( \theta_X(s) = \theta_X(t) \) for \( s \in [t, t + \epsilon] \) (right-continuity). See Figure 2 for an example.

\[\]

Figure 2: A treegram \( \theta_X \) over the set \( X = \{x_1, x_2, x_3, x_4\} \). Notice that \( \theta_X(t) = \emptyset \) for \( t \in (-\infty, t_0) \). Also, \( \theta_X(t_0) = \{\{x_1\}\} \), \( \theta_X(t_2) = \{\{x_1\}, \{x_2, x_3\}\} \), and \( \theta_X(t) = \{X\} \) for all \( t \in [t_3, \infty) \).

2.2 A distance between dendrograms

In this section we review the method of \cite{7} for quantifying the structural difference between dendrograms. In short, we compare two dendrograms over sets \( X \) and \( Y \) by comparing their associated ultrametrics on \( X \) and \( Y \), respectively.

Dendrograms and their associated ultrametrics. An ultrametric space \( (X, u_X) \) is a metric space satisfying the strong triangle inequality: for all \( x, x', x'' \in X \), \( u_X(x, x') \leq \max\{u_X(x, x''), u_X(x'', x')\} \).

Let \( X \) be a finite set and let \( \theta_X : R_+ \to \mathcal{P}(X) \) be a dendrogram over \( X \). Recall from \cite{7} that this \( \theta_X \) induces a canonical ultrametric \( u_{\theta_X} : X \times X \to R_+ \) on \( X \) defined by
\[
u_{\theta_X}(x, x') := \inf\{\epsilon \geq 0 : x, x' \text{ belong to the same block of } \theta_X(\epsilon)\}.
\]

For example, for the dendrogram \( \theta_X \) depicted in Figure 1 it is easy to observe that \( u_{\theta_X}(x_1, x_4) = T \).

Reciprocally, any ultrametric space \( (X, u_X) \) induces a dendrogram \( \theta_X \) over \( X \) \cite{7}.

The Gromov-Hausdorff distance \cite{4, 7}. The Gromov-Hausdorff distance quantifies how far two compact metric spaces are from being isometric. This distance is widely used in applications such as shape comparison (for example, see \cite{20}). In order to define the

\[\]

\footnote{In order to regard a dendrogram \( \theta_X : R_+ \to \mathcal{P}(X) \) as a treegram, trivially extend \( \theta_X \) to the whole \( R_+ \) for \( t \in (-\infty, 0) \), let \( \theta_X(t) := \emptyset \in \mathcal{P}_{\text{sub}}(X) \) by definition.}
Gromov-Hausdorff distance, one needs the notion of correspondence. For sets \(X\) and \(Y\), a subset \(R \subset X \times Y\) is said to be a correspondence (between \(X\) and \(Y\)) if and only if (1) for every \(x \in X\), there exists \(y \in Y\) such that \((x, y) \in R\), and (2) for every \(y \in Y\), there exists \(x \in X\) such that \((x, y) \in R\).

Let \((X, d_X)\) and \((Y, d_Y)\) be any two compact metric spaces. The Gromov-Hausdorff distance between \((X, d_X)\) and \((Y, d_Y)\) is defined by

\[
d_{GH}((X, d_X), (Y, d_Y)) := \frac{1}{2} \inf_R \sup_{(x,y) \in R} |d_X(x, x') - d_Y(y, y')|,
\]

where the infimum is taken over all correspondences between \(X\) and \(Y\). Note that in the case where \((X, d_X)\) and \((Y, d_Y)\) are finite metric spaces, the infimum and the supremum above can be replaced with the minimum and the maximum, respectively.

**A distance between dendrograms.** Let \(\theta_X\) and \(\theta_Y\) be dendrograms over finite sets \(X\) and \(Y\), respectively. One defines the Gromov-Hausdorff distance \([7]\) between the dendrograms \(\theta_X\) and \(\theta_Y\) as

\[
d_{GH}(\theta_X, \theta_Y) := d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y})),
\]

where \(u_{\theta_X}\) and \(u_{\theta_Y}\) are the ultrametrics associated to the dendrograms \(\theta_X\) and \(\theta_Y\), respectively.

### 2.3 Finest common coarsening of (sub-)partitions

For a set \(X\), we know that there exists a canonical one-to-one correspondence between the collection of all equivalence relations on \(X\) and the collection of all partitions \(\mathcal{P}(X)\) of \(X\). We will extend this correspondence in a certain way for defining the notion of finest common coarsening in the collection \(\mathcal{P}^{\text{sub}}(X)\) of all sub-partitions of \(X\).

**Sub-equivalence relations.** Let \(X\) be a non-empty set. Let \(\sim\) be any equivalence relation on any subset \(X' \subset X\). We call the relation \(\sim\) a sub-equivalence relation on \(X\). We also call \(X'\) the underlying set of \(\sim\), which is identical to \(\{x \in X : (x, x) \in \sim\}\).

Clearly, any equivalence relation on \(X\) is also a sub-equivalence relation with underlying set \(X\).

There is the canonical one-to-one correspondence between the collection of all sub-equivalence relations on \(X\) and the collection \(\mathcal{P}^{\text{sub}}(X)\) of all sub-partitions of \(X\): Any sub-equivalence relation \(\sim\) on \(X\) corresponds to the sub-partition \(P\) with underlying set \(X' = \{x \in X : (x, x) \in \sim\}\) such that \(x \sim y\) if and only if \(x\) and \(y\) belong to the same block \(B \in P\). Reciprocally, to any sub-partition \(P\) of \(X\), one can associate the unique sub-equivalence relation \(\sim_P\) on \(X\) defined by \(x \sim_P y\) if and only if \(x\) and \(y\) belong to the same block \(B \in P\).

**Sub-equivalence closure.** Let \(X\) be a non-empty set. For an index set \(I\), suppose that \(\{\sim_i \subset X \times X : i \in I\}\) is a collection of sub-equivalence relations on \(X\). The sub-equivalence closure of the collection \(\{\sim_i \subset X \times X : i \in I\}\) is defined to be the transitive closure of the relation \(\bigcup_{i \in I} \sim_i\) on \(X\). In other words, by the sub-equivalence closure of the collection \(\{\sim_i \subset X \times X : i \in I\}\), we mean the minimal sub-equivalence relation containing \(\sim_i\) for all \(i \in I\).

**Finest common coarsening.** Let \(\{P_i\}_{i \in I}\) be any sub-collection of \(\mathcal{P}^{\text{sub}}(X)\). For each \(i \in I\), let \(\sim_i\) be the sub-equivalence relation on \(X\) corresponding to \(P_i\). By \(\bigvee_{i \in I} P_i\), we mean the sub-partition of \(X\) corresponding to the sub-equivalence closure of the collection \(\{\sim_i \subset X \times X : i \in I\}\). We will refer to \(\bigvee_{i \in I} P_i\) as the finest common coarsening of the collection \(\{P_i\}_{i \in I}\).

For example, let \(X = \{x, y, z, w\}\). For \(P_1 = \{\{x\}, \{y\}\}\), \(P_2 = \{\{y, z\}\}\), and \(P_3 = \{\{x, w\}\}\) in \(\mathcal{P}^{\text{sub}}(X)\), we have:

1. \(\bigvee_{i=1}^2 P_i = \{\{x, y, z\}\} \in \mathcal{P}^{\text{sub}}(X)\), and
2. \(\bigvee_{i=1}^3 P_i = \{\{x, w\}, \{y, z\}\} \in \mathcal{P}(X)\).

### 3 Formigrams

Although the notions of dendrogram or treegram are useful when representing the output of a hierarchical clustering method (i.e. when partitions only become coarser with the increase of a parameter), in order to represent the diverse clustering behaviors of dynamic datasets we need a more flexible concept allowing for possible refinement of partitions. Here we suggest a “zigzag like” notion of dendrograms that we call formigram. We allow partitions to become finer sometimes, but require that partitions defined by a formigram change only finitely many times in any finite interval for visualization.

#### 3.1 The definition of a formigram

**Formigrams.** A formigram over a finite set \(X\) is any function \(\theta_X : \mathbb{R} \to \mathcal{P}^{\text{sub}}(X)\) such that:

1. (Tameness) the set \(\text{crit}(\theta_X)\) of points of discontinuity of \(\theta_X\) is locally finite[4]. We call the elements of \(\text{crit}(\theta_X)\) the critical points of \(\theta_X\).

[4]To say that \(\text{crit}(\theta_X)\) is locally finite means that for any bounded interval \(I \subset \mathbb{R}\), the cardinality of \(I \cap \text{crit}(\theta_X)\) is finite. The purpose of this condition is twofold: on the one hand,
2. (Interval lifespan) for every \( x \in X \), the set \( I_x := \{ t \in \mathbb{R} : x \in B \in \theta_X(t) \} \), said to be the lifespan of \( x \), is a non-empty closed interval.

3. (Comparability) for every point \( c \in \mathbb{R} \) it holds that
\[
\theta_X(c - \varepsilon) \subseteq \theta_X(c) \subseteq \theta_X(c + \varepsilon)
\]
for all sufficiently small \( \varepsilon > 0 \).

Note that the definition of formigrams generalizes those of dendrograms and treegrams. In other words, every dendrogram and every treegram are formigrams. See Figure [3] for an example.

\[
\theta_X(t) = \begin{cases} 
\{\{x_1, x_2, x_3\}\}, & t \in (-\infty, t_0) \\
\{\{x_1, x_2, x_3\}, \{x_4\}\}, & t \in [t_0, t_1) \\
\{\{x_1, x_2, x_3, x_4\}\}, & t \in [t_1, t_2] \cup [t_2, \infty) \\
\{\{x_2, x_3, x_4\}\}, & t \in (t_2, t_3] \cup [t_4, t_5) \\
\{\{x_1\}\}, & t \in (t_3, t_4) \\
\end{cases}
\]

Figure 3: Top: The specification of a formigram \( \theta_X \) over the set \( X = \{x_1, x_2, x_3, x_4\} \). Bottom: A graphical representation of the formigram \( \theta_X \).

### 3.2 A distance between formigrams

In this section we introduce a (pseudo) metric on the collection of all formigrams. This metric quantifies the structural difference between two grouping/disbanding behaviors over time. In particular, when restricting this metric to the collection of dendrograms, (twice) the Gromov-Hausdorff distance between dendrograms is recovered (Theorem 2).

#### Partition morphisms.
Before introducing a metric on formigrams, we first establish a method for interconnecting any two partitions with possibly different underlying sets. Recall that for any sets \( X \) and \( Y \), a multivalued map \( \varphi : X \rightrightarrows Y \) is a relation between \( X \) and \( Y \)

such that for all \( x \in X \), there exists (a not necessarily unique) \( y \in Y \) with \( (x, y) \in \varphi \). For \( x \in X \), the image \( \varphi(x) \) of \( x \) is defined to be the set \( \{y \in Y : (x, y) \in \varphi\} \).

For any two sets \( X \) and \( Y \), let \( P_X \in \mathcal{P}(X) \) and \( P_Y \in \mathcal{P}(Y) \). Any multivalued map \( \varphi : X \rightrightarrows Y \) (or map \( \varphi : X \rightarrow Y \)) is said to be a partition morphism between \( P_X \) and \( P_Y \) if for any \( x, x' \in X \) belonging to the same block of \( P_X \), their images \( \varphi(x), \varphi(x') \) are included in the same block of \( P_Y \) (note that \( \varphi(x), \varphi(x') \) can be sets containing more than one element). In this case, we write \( P_X \leq_Y P_Y \).

If \( P_X \leq_Y P_Y \), then there exists the canonical induced map \( \varphi^* : P_X \rightarrow P_Y \) defined by sending each block \( B \in P_X \) to the block \( C \in P_Y \) such that \( \varphi(B) \subset C \).

**A distance between formigrams.** Exploiting the fact that any formigram is a “stack” of (sub-)partitions of a specific set, we now introduce the interleaving distance \( d_I^\varepsilon \) on the collection of all formigrams. The construction of \( d_I^\varepsilon \) is inspired by the interleaving distance for Reeb graphs.

Let \( \theta_X \) be a formigram over \( X \) and let \( I \subset \mathbb{R} \) be an interval. We define \( \bigvee_{t \in I} \theta_X \) to be the finest common coarsening of the collection \( \{\theta_X(t) : t \in I\} \) of sub-partitions of \( X \). Also, for any \( t \in \mathbb{R} \), define \( [t]^\varepsilon := [t - \varepsilon, t + \varepsilon] \subset \mathbb{R} \).

Let \( \theta_X \) and \( \theta_Y \) be any two formigrams over \( X \) and \( Y \), respectively. \( \theta_X \) and \( \theta_Y \) are said to be \( \varepsilon \)-interleaved if there exists a correspondence \( R \) between \( X \) and \( Y \) satisfying the following:

1. For any \( (x, y) \in R \) and any \( t \in \mathbb{R} \),
   (a) if \( x \) is in the underlying set of \( \theta_X(t) \), then \( y \) is in the underlying set of \( \bigvee_{t \in [t]^\varepsilon} \theta_Y \)
   (b) if \( y \) is in the underlying set of \( \theta_Y(t) \), then \( x \) is in the underlying set of \( \bigvee_{t \in [t]^\varepsilon} \theta_X \).

2. For all \( t \in \mathbb{R} \),
   \[
   \theta_X(t) \leq_R \bigvee_{|t|^\varepsilon} \theta_Y \quad \text{and} \quad \theta_Y(t) \leq_{R^{-1}} \bigvee_{|t|^\varepsilon} \theta_X,
   \]
   where \( R^{-1} = \{(y, x) \in Y \times X : (x, y) \in R\} \).

We call any such \( R \) an \( \varepsilon \)-correspondence between \( \theta_X \) and \( \theta_Y \). The interleaving distance \( d_I^\varepsilon (\theta_X, \theta_Y) \) between \( \theta_X \) and \( \theta_Y \) is defined by the infimum of \( \varepsilon \geq 0 \) for which there exists an \( \varepsilon \)-correspondence between \( \theta_X \) and \( \theta_Y \). If there is no \( \varepsilon \)-correspondence between \( \theta_X \) and \( \theta_Y \) for any \( \varepsilon \geq 0 \), then we declare \( d_I^\varepsilon (\theta_X, \theta_Y) = +\infty \).

---

\[\text{In particular, any correspondence } R \text{ between } X \text{ and } Y \text{ is a multivalued map.}\]

\[\text{We remark that this condition is equivalent to saying that if} \]

\[\text{if } x \text{ is in the underlying set of } \theta_X(t), \text{ then there exists } t_0 \in [t]^\varepsilon \text{ such that} \]

\[\text{if } y \text{ is in the underlying set of } \theta_Y(t_0).\]

\[\text{Note that if } R \text{ is an } \varepsilon \text{-correspondence between } \theta_X \text{ and } \theta_Y, \text{ then for any } \varepsilon' > \varepsilon, \text{ } R \text{ is also an } \varepsilon' \text{-correspondence between } \theta_X \text{ and } \theta_Y.\]
Theorem 1 $d^F_1$ is an extended pseudo-metric on formigrams.

See Appendix A for the proof of Theorem 1. For example, consider any formigram $\theta_X$ over a finite set $X$ and let $t > 0$. Define another formigram $\theta_X^t$ as $\theta_X^t(t) := \theta_X(t + \tau)$ for $t \in \mathbb{R}$. Then, it is not difficult to verify that $d^F_1$($\theta_X, \theta_X^t$) $\leq\tau$ by checking that $R_X := \{(x, x) : x \in X\}$ is a $\tau$-correspondence between $\theta_X$ and $\theta_X^t$.

Theorem 2 $d^F_1$ generalizes the Gromov-Hausdorff distance between dendrograms. Namely, for any dendrograms $\theta_X$ and $\theta_Y$ over $X$ and $Y$ respectively,

$$d^F_1(\theta_X, \theta_Y) = 2d_{GH}(\theta_X, \theta_Y).$$

Proof. Recall that by definition

$$d_{GH}(\theta_X, \theta_Y) = d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y}))$$

where $u_{\theta_X}$ and $u_{\theta_Y}$ are the ultrametrics associated to the dendrograms $\theta_X$ and $\theta_Y$ respectively. Therefore, we will show that $d^F_1(\theta_X, \theta_Y) = 2d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y}))$.

First we show “$\leq$”. If $d^F_1(\theta_X, \theta_Y) = \infty$, there is nothing to prove and hence we assume that $d^F_1(\theta_X, \theta_Y)$ is finite. Then, there exists an $\varepsilon$-correspondence $R \subset X \times Y$ between the two dendrograms $\theta_X$ and $\theta_Y$ for some $\varepsilon \geq 0$, implying that $d^F_1(\theta_X, \theta_Y) \leq \varepsilon$. Pick any $(x, y), (x', y') \in R$ and let $t := u_{\theta_X}(x, x')$. Then, $x, x'$ belong to the same block of the partition $\theta_X(t)$. Since $\theta_X(t) \leq \bigvee_{t'} \theta_Y$, $y, y'$ must belong to the same block of $\bigvee_{t'} \theta_Y$. Also, since $\theta_Y$ is a dendrogram, $\theta_Y(s_1) \leq \theta_Y(s_2)$ for any $s_1 \leq s_2$, and thus $\bigvee_{t'} \theta_Y = \theta_Y(t + \varepsilon)$. Therefore, $y, y'$ belong to the same block of $\theta_Y(t + \varepsilon)$, and in turn $u_{\theta_Y}(y, y') \leq t + \varepsilon = u_{\theta_X}(x, x') + \varepsilon$. By symmetry, we also have $u_{\theta_X}(x, x') \leq u_{\theta_Y}(y, y') + \varepsilon$. Therefore, by the definition of $d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y}))$, we have

$$d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y})) \leq \varepsilon/2.$$

Next, we prove “$\geq$”. Let $R$ be a correspondence between $X$ and $Y$ such that for all $(x, y), (x', y') \in R$, $|u_{\theta_X}(x, x') - u_{\theta_Y}(y, y')| \leq \varepsilon$, implying that $d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y})) \leq \varepsilon/2$. We wish to show that $\theta_X(t) \leq \bigvee_{t'} \theta_Y(t + \varepsilon)$ for all $t \in \mathbb{R}$. For $t < 0$, since $\theta_X(t) = \theta_Y(t) = 0$, we trivially have $\theta_X(t) \leq \bigvee_{t'} \theta_Y(t + \varepsilon)$. Now pick any $t \geq 0$ and any $(x, y), (x', y') \in R$. Assume that $x, x'$ belong to the same block of $\theta_X(t)$, implying that $u_{\theta_X}(x, x') \leq t$. Since $|u_{\theta_X}(x, x') - u_{\theta_Y}(y, y')| \leq \varepsilon$, we know $u_{\theta_Y}(y, y') \leq t + \varepsilon$ and hence $y, y'$ belong to the same block of $\bigvee_{t'} \theta_Y(t + \varepsilon)$. Therefore, $\theta_X(t) \leq \bigvee_{t'} \theta_Y(t + \varepsilon)$ for all $t \in \mathbb{R}$. By symmetry, $\theta_Y(t) \leq \bigvee_{t'} \theta_Y(t + \varepsilon)$ for all $t \in \mathbb{R}$ as well, completing the proof.

Theorem 3 (Complexity of computing $d^F_1$) Fix $\rho \in (1, 6)$. It is not possible to obtain a $\rho$ approximation to the distance $d^F_1((X, \theta_X), (Y, \theta_Y))$ between formigrams in time polynomial on $|X|, |Y|, \text{crit}(\theta_X), \text{crit}(\theta_Y)$ unless $P = NP$.

Proof. Pick any two dendrograms $\theta_X$ and $\theta_Y$ and invoke Theorem 1 to reduce the problem to the computation of the Gromov-Hausdorff distance

$$\Delta := d_{GH}((X, u_{\theta_X}), (Y, u_{\theta_Y}))$$

between the ultrametric spaces $(X, u_{\theta_X}), (Y, u_{\theta_Y})$ associated to the dendrograms. However, according to [22, Corollary 3.8], $\Delta$ cannot be approximated within any factor less than 3 in polynomial time, unless $P = NP$. The author shows this by observing that any instance of the 3-partition problem can be reduced to an instance of the bottleneck $\infty$-Gromov-Hausdorff distance ($\infty$-BGHD) problem between ultrametric spaces (see [22, p.865]). The proof follows.

4 Application: Visualization of clustering features of dynamic metric data

In this section we explain how to extract scale dependent clustering features from time-dependent metric spaces in the form of formigrams. Furthermore, we will show that this summarization process is stable under perturbations in the input time-dependent metric spaces.

4.1 Dynamic metric spaces (DMSs)

Recall that a pseudo-metric space is a pair $(X, d_X)$ where $X$ is a (non-empty) set and $d_X : X \times X \to \mathbb{R}_+$ is a symmetric function which satisfies the triangle inequality, and such that $d_X(x, x) = 0$ for all $x \in X$. $d_X$ is called the pseudo-metric. Note that one does not require $d_X(x, x') = 0$ implies that $x = x'$ like in the case of standard metric spaces.

Dynamic metric spaces (DMSs). A dynamic metric space is a pair $\gamma_X = (X, d_X(\cdot))$ where $X$ is a non-empty finite set and $d_X : \mathbb{R} \times X \times X \to \mathbb{R}_+$ is a symmetric function which satisfies the triangle inequality, and such that $d_X(x, x) = 0$ for all $x \in X$. $d_X$ is called the pseudo-metric. Note that one does not require $d_X(x, x') = 0$ implies that $x = x'$ like in the case of standard metric spaces.

Dynamic metric spaces (DMSs). A dynamic metric space is a pair $\gamma_X = (X, d_X(\cdot))$ where $X$ is a non-empty finite set and $d_X : \mathbb{R} \times X \times X \to \mathbb{R}_+$ satisfies:

1. For every $t \in \mathbb{R}$, $\gamma_X(t) = (X, d_X(t))$ is a pseudo-metric space.
2. There exists $t_0 \in \mathbb{R}$ such that $\gamma_X(t_0)$ is a (standard) metric space.
3. For fixed $x, x' \in X$, $d_X(\cdot)(x, x') : \mathbb{R} \to \mathbb{R}_+$ is continuous.

We refer to $t$ as the time parameter. Condition 2 above is assumed since otherwise one could substitute the DMSs $\gamma_X$ by another DMSs $\gamma_{X'}$ over a set $X'$ which satisfies $|X'| < |X|$, and such that $\gamma_{X'}$ is point-wise equivalent to $\gamma_X$.

A family of examples of DMSs is given by $n$ particles/animals moving continuously inside an environment $\Omega \subset \mathbb{R}^d$ where particles are allowed to coalesce. If the $n$ trajectories are $p_1(t), \ldots, p_n(t) \in \mathbb{R}^d$, then let
Appendix A. See Figure 4 for an illustration. We prove Theorem 4 in Appendix A. Then, the function $C_δ(γ_X) : R \rightarrow P(X)$ defined by $C_δ(γ_X)(t) = C_δ(γ_X(t))$ for $t ∈ R$ is a formigram.

See Figure 4 for an illustration. We prove Theorem 4 in Appendix A.

4.3 Stability of $δ$-clustering method for DMSs.

It turns out that the construction of formigrams from DMSs described in Theorem 4 is stable under perturbations in the input DMSs under a certain notion of distance between DMSs described below. Structurally, this distance is a hybrid between the Gromov-Hausdorff distance and the interleaving distance \[2, 8\] for Reeb graphs \[9\].

A distance between DMSs. Let $γ_X, γ_Y$ be DMSs and $ε ≥ 0$. We say that $γ_X$ and $γ_Y$ are $ε$-interleaved if there exists a correspondence $R$ between $X$ and $Y$ such that (*): $(∀(x, y), (x', y') ∈ R, ∀t ∈ R,$

1. $\min_{s ∈ [t]} d_Y(s)(y, y') ≤ d_X(t)(x, x')$ and,
2. $\min_{s ∈ [t]} d_X(s)(x, x') ≤ d_Y(t)(y, y')$.

When $γ_X$ and $γ_Y$ are $ε$-interleaved we write $γ_X ≈_ε γ_Y$. The interleaving distance between $γ_X$ and $γ_Y$ is defined by $d^\text{dyn}_ε(γ_X, γ_Y) := \inf_{ε ≥ 0} (γ_X ≈_ε γ_Y)$. If $γ_X$ and $γ_Y$ are not $ε$-interleaved for any $ε ≥ 0$, declare $d^\text{dyn}_ε(γ_X, γ_Y) = +∞$. Also, any correspondence $R$ satisfying (*) is called an $ε$-correspondence between $γ_X$ and $γ_Y$.

In Appendix B, we show that $d^\text{dyn}_ε$ is indeed an extended metric on DMSs (Theorem 5).

Theorem 5 (Stability theorem) For any tame DMSs $γ_X, γ_Y$ and any $δ ≥ 0$, let $θ_X := C_δ(γ_X)$ and $θ_Y := C_δ(γ_Y)$ as in Theorem 4. Then,

$d^θ_δ(θ_X, θ_Y) ≤ d^\text{dyn}_δ(γ_X, γ_Y)$.

See Appendix B for the proof of Theorem 5.

5 Conclusion and Discussion

We introduced formigrams: a generalization of the notion of dendrograms that is useful for characterizing and visualizing the clustering features of DMSs. We clarified a sufficient condition (tameness) for DMSs to admit a summarization as formigrams.

We also defined the distances $d^θ_δ$ and $d^\text{dyn}_δ$ on formigrams and on DMSs, respectively, and showed that the $δ$-clustering method for DMSs is stable under perturbations in the input DMSs in terms of $d^θ_δ$ and $d^\text{dyn}_δ$. Specifically, it is noteworthy that $d^θ_δ$ generalizes the Gromov-Hausdorff distance on dendrograms.

In [17], due to the high cost of computing $d^θ_δ$, we carry out a classification task for different flocking behaviors by making use of a tractable lower bound for $d^θ_δ$. The nature of this lower bound is related to zigzag persistence theory [5, 6]: One can find theoretical details in [15].
References


Appendix A

Proof of Theorem 4.

Proof. We show that $\theta_X := C_4(\gamma_X)$ satisfies the three conditions (tameness, interval lifespan, and comparability) to be a formigram. First, by the definition of $C_3$, $C_4(\gamma_X)$ is a function from $R$ to the set of all partitions $\mathcal{P}(X) \subset \mathcal{P}_{\text{sub}}(X)$ of $X$. Therefore, every element $x \in X$ has the full lifespan $l_x = (\infty, \infty)$, in $\theta_X$.

Next we show the comparability condition. For simplicity, assume that $X = \{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$. Fix $c \in R$ and consider the following two subsets of $X \times X$:

$$A(c, \delta) := \{(i,j) : i < j \in X, \ d_X(i,j) \leq \delta\},$$

$$B(c, \delta) := \{(i,j) : i < j \in X, \ d_X(i,j) > \delta\}.$$

The continuity of $d_X(\cdot)(i,j)$ for each $(i,j) \in X \times X$ guarantees that there exists $\varepsilon > 0$ such that

$$B(t, \delta) \supseteq B(c, \delta) \quad \text{for all } t \in (c-\varepsilon, c+\varepsilon)$$

and in turn

$$A(t, \delta) \subseteq A(c, \delta) \quad \text{for all } t \in (c-\varepsilon, c+\varepsilon).$$

Since $A(t, \delta) \cup B(t, \delta) = \{(i,j) : i < j \in X\}$ for all $t \in R$. This implies that the partition $C_3(\gamma_X(c))$ is coarser than or equal to $C_4(\gamma_X(t))$ for each $t \in (c-\varepsilon, c+\varepsilon)$, which means that $C_4(\gamma_X)$ satisfies the comparability condition.

It remains to prove that $C_4(\gamma_X)$ is tame. For $i, j \in X$, let $f_{i,j} := d_X(\cdot)(i,j) : R \to R_1$ and let $I \subseteq R$ be any finite interval. Note that discontinuity points of the function $C_4(\gamma_X) : R \to \mathcal{P}(X)$ can occur only at endpoints of connected components of the set $f_{i,j}^{-1}(\delta)$ for some $i, j \in X$. Fix any $i, j \in X$. Then, since $\gamma_X$ is tame, the set $f_{i,j}^{-1}(\delta) \cap I$ has only finitely many connected components and thus there are only finitely many endpoints arising from those components.

Since the set $X$ is finite, this implies that $C_4(\gamma_X)$ can have only finitely many critical points in $I$. \hfill \square

Appendix B

Isomorphic DMSs. We now introduce a notion of equality between two DMSs. Let $\gamma_X = (X, d_X(\cdot))$ and $\gamma_Y = (Y, d_Y(\cdot))$ be DMSs. We say that $\gamma_X$ and $\gamma_Y$ are isomorphic if there exists a bijection $\varphi : X \to Y$ such that $\varphi$ is an isometry between $\gamma_X(\cdot)$ and $\gamma_Y(\cdot)$ across all $t \in R$.

Theorem 6. $d_X^{\text{syn}}$ is an extended metric modulo isomorphisms between DMSs.

We will prove Theorem 6 after showing Theorem 5.

Proof of Theorem 5.

Proof. First, note that for all $t \in R$, $X$ and $Y$ are the underlying sets of $\theta_X(\cdot)$ and $\theta_Y(\cdot)$, respectively.

Let $\varepsilon \geq 0$ and assume that $R \subseteq X \times Y$ is any $\varepsilon$-correspondence between $\gamma_X$ and $\gamma_Y$. It suffices to prove that $R$ is an $\varepsilon$-correspondence between the formigrams $\theta_X$ and $\theta_Y$ to the same block of $V_{[t]^2} \theta_Z$. Similarly, one can verify that $\theta_Z(t) \leq (R_2 \circ R_1)^{-1} V_{[t]^2} \theta_X$. \hfill \square
as well. Let \((x, y), (x', y') \in R\) and fix any \(t \in \mathbb{R}\). Assume that \(x, x'\) belong to the same block of \(\theta_X(t)\), meaning that there is a sequence \(x = x_0, x_1, \ldots, x_n = x'\) in \(X\) such that 
\[d_X(t)(x_i, x_{i+1}) \leq \delta \text{ for } 0 \leq i \leq n - 1.\]
For each \(0 \leq i \leq n - 1\), pick \(y_i \in Y\) such that \((x_i, y_i) \in R\) where \(y = y_0\) and \(y' = y_n\). Since \(R\) is an \(\varepsilon\)-correspondence between \(\gamma_X, \gamma_Y\), we have
\[\min_{s \in [t]} d_Y(s)(y_i, y_{i+1}) \leq d_X(t)(x_i, x_{i+1}) \leq \delta.\]
This implies that, for each \(i\), there is \(s_i \in [t]\) such that \(d_y(s_i)(y_i, y_{i+1}) \leq \delta\) and in turn \(y_i, y_{i+1}\) are in the same block of \(\theta_Y(s_i)\). Also for each \(i\), since \(s_i \in [t]\), one has \(\theta_Y(s_i) \leq \sqrt{[t]} \theta_Y\) and in turn \(y_i, y_{i+1}\) belong to the same block of \(\sqrt{[t]} \theta_X\). Therefore, we conclude that \(y, y'\) belong to the same block of \(\sqrt{[t]} \theta_X\). We have proved that \(\theta_X(t) \leq R \sqrt{[t]} \theta_Y\). Similarly, \(\theta_Y(t) \leq R^{-1} \sqrt{[t]} \theta_X\) can be shown, completing the proof. \(\square\)

**Proof of Theorem 6.**

Proof. Reflexivity and symmetry of \(d_1^{\text{dyn}}\) are clear so we shall show the triangle inequality only: that for all DMSs \(\gamma_X, \gamma_Y, \gamma_Z\), one has \(d_1^{\text{dyn}}(\gamma_X, \gamma_Z) \leq d_1^{\text{dyn}}(\gamma_X, \gamma_Y) + d_1^{\text{dyn}}(\gamma_Y, \gamma_Z)\). We assume that \(d_1^{\text{dyn}}(\gamma_X, \gamma_Y)\) and \(d_1^{\text{dyn}}(\gamma_Y, \gamma_Z)\) are finite because otherwise there is nothing to prove. Let \(0 < \varepsilon_1, \varepsilon_2 < \infty\) ans suppose that there are an \(\varepsilon_1\)-correspondence \(R_1 \subset X \times Y\) between \(\gamma_X\) and \(\gamma_Y\) and an \(\varepsilon_2\)-correspondence \(R_2 \subset Y \times Z\) between \(\gamma_Y\) and \(\gamma_Z\). Define the correspondence \(R_2 \circ R_1\) between \(X\) and \(Z\) as follows:

\[R_2 \circ R_1 := \{(x, z) \in X \times Z : \exists y \in Y \text{ s.t. } (x, y) \in R_1 \text{ and } (y, z) \in R_2\}.\]

Pick any two pairs \((x, z)\) and \((x', z')\) in \(R_2 \circ R_1\). Then, there are \(y, y' \in Y\) such that \((x, y), (x', y') \in R_1\) and \((y, z), (y', z') \in R_2\). Then for all \(t \in \mathbb{R}\), it holds that

\[\min_{s \in [t]} d_Z(s)(z, z') \leq \min_{s \in [t]} d_Y(s)(y, y') \leq d_X(t)(x, x'),\]
\[\min_{s \in [t]} d_X(s)(x, x') \leq \min_{s \in [t]} d_Y(s)(y, y') \leq d_Z(t)(z, z').\]

Therefore, \(R_2 \circ R_1\) is an \((\varepsilon_1 + \varepsilon_2)\)-correspondence between \(\gamma_X, \gamma_Z\), implying that
\[d_1^{\text{dyn}}(\gamma_X, \gamma_Z) \leq d_1^{\text{dyn}}(\gamma_X, \gamma_Y) + d_1^{\text{dyn}}(\gamma_Y, \gamma_Z),\] as desired.

Now, we show that \(d_1^{\text{dyn}}\) is not just an (extended) pseudo-metric but an (extended) metric. Assume that \(d_1^{\text{dyn}}(\gamma_X, \gamma_Y) = 0\) for some DMSs \(\gamma_X, \gamma_Y\). Since there exist only finitely many correspondences between \(X\) and \(Y\), there must exist a correspondence \(R \subset X \times Y\) such that for any \(\varepsilon > 0\), \(R\) is an \(\varepsilon\)-correspondence between \(\gamma_X\) and \(\gamma_Y\). We claim that this \(R\) is a 0-correspondence. To this end, we need the following:

**Claim.** Let \(f : \mathbb{R} \to \mathbb{R}\) be a continuous map and \(r, t \in \mathbb{R}\). Suppose that for every \(\varepsilon > 0\), \(\min_{s \in [t]} f(s) \leq r\). Then \(f(t) \leq r\).

**Proof.** [Proof of Claim] For each \(k \in \mathbb{N}\), take any \(s_k \in [t]^{1/k}\) such that \(f(s_k) \leq r\). Then \((s_k)_{k \in \mathbb{N}}\) is a sequence in \(f^{-1}(-\infty, r)\) converging to \(t\). Since \(f\) is continuous, \(f^{-1}(-\infty, r)\) is a closed set and thus \(t\) must belong to \(f^{-1}(-\infty, r)\), i.e. \(f(t) \leq r\), as desired. \(\square\)