A note on maximally repeated sub-patterns of a point set

Véronique Cortier, Xavier Goaoc, Mira Lee, Hyeon-Suk Na

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A note on maximally repeated sub-patterns of a point set

Véronique Cortier — Xavier Goaoc — Mira Lee — Hyeon-Suk Na

N° 5773

Decembre 2005

Thème SYM

ISSN 0249-6399 ISRN INRIA/RR--5773--FR+ENG

A note on maximally repeated sub-patterns of a point set

Véronique Cortier* , Xavier Goaoc† , Mira Lee‡ , Hyeon-Suk Na§

Thème SYM — Systèmes symboliques
Projets Cassis et Vegas

Rapport de recherche n° 5773 — Decembre 2005 — 5 pages

Abstract: We answer a question raised by P. Brass on the number of maximally repeated sub-patterns in a set of n points in $\mathbb{R}^d$. We show that this number, which was conjectured to be polynomial, is in fact $\Theta(2^{n/2})$ in the worst case, regardless of the dimension d.

Key-words: Discrete geometry, point sets, repeated configurations.

* Projet Cassis, LORIA - CNRS, Nancy, France. Email: cortier@loria.fr.
† Projet Vegas, LORIA - INRIA Lorraine, Nancy, France. Email: goaoc@loria.fr.
‡ Division of Computer Science, Korea Advanced Institure of Science and Technology (KAIST), Republic of Korea. Email: mira@kaist.ac.kr
§ School of Computing, Soongsil University, Seoul, Republic of Korea. Email: hsnaa@computing.ssu.ac.kr

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,

Une note sur les sous-motifs maximallement répétés d'un nuage de points

Résumé : Nous répondons à une question de P. Brass sur le nombre de sous-motifs maximallement répétés d'un ensemble de n points de $\mathbb{R}^d$. Nous montrons que ce nombre, conjecturé polynomial, s'avère être $\Theta(2^{n/2})$ dans le cas le pire, et ce en toute dimension d.

Mots-clés : Géométrie discrète, nuages de points, sous-motifs répétés.

1 Introduction

Let $\mathcal{S}$ be a set of n points in $\mathbb{R}^d$. A *sub-pattern*, i.e. a subset, of $\mathcal{S}$ is repeated if it can be translated to another subset of $\mathcal{S}$. A sub-pattern $P \subseteq \mathcal{S}$ is *maximally repeated* if for any subset Q such that $P \subsetneq Q \subseteq \mathcal{S}$ there exists a translation that maps P to a subset of $\mathcal{S}$ without mapping Q to a subset of $\mathcal{S}$. In other words, a pattern is maximally repeated if it cannot be extended without losing at least one of its occurrences. Maximally repeated sub-patterns (MRSP for short) originated from the field of pattern matching to solve the following problem: given two point sets X and Y, can Y be translated to a subset of X? P.Brass [1, Theorem3] gave an algorithm that answers such queries in time $O(|Y| \log |X|)$ whose preprocessing time depends on the number of distinct MRSP of X, where two MRSP are *distinct* if they are not equal up to a translation. A natural question is thus to give a theoretical bound on this number of MRSP in order to provide an upper bound on the time requirement of that algorithm. This number was conjectured [1] [2, p.267] to be $O(n^d)$ where d is the dimension in which the point set is embedded.

In this note we show that the number of MRSP of a set of n points is actually $\Theta(2^{n/2})$ in the worst case, which shows that finding sub-patterns via this approach may lead to exponential worst-case running time. Our proof is based on combinatorial rather than geometrical properties of the point set, which explains that the bound is independent of the dimension d in which the points are considered.

2 Lower and upper bounds

Let us first introduce some terminology. Given a set of points $P \subseteq \mathbb{R}^d$ and a translation $t \in \mathbb{R}^d$, $P+t := \{x+t \mid x \in P\}$ is the set of translated points of P by t. A subset $P \subseteq \mathcal{S}$ is a repeated sub-pattern if there exists a translation $t \neq \mathbf{0}$ such that $P+t \subseteq \mathcal{S}$. P is a *maximally repeated sub-pattern* (MRSP) if, in addition, for any subset Q such that $P \subsetneq Q \subseteq \mathcal{S}$ there exists a translation t such that $P + t \subseteq \mathcal{S}$ and $Q + t \nsubseteq \mathcal{S}$. Two MRSP are *distinct* if they are not equal up to a translation.

In the sequel, we present a set of n points in $\mathbb{R}$ having at least $2^{\lfloor n/2 \rfloor - 1}$ distinct MRSP (Section 2.1) and then prove that any set of n points in $\mathbb{R}^d$ can have at most $16 \cdot 2^{\lceil n/2 \rceil}$ distinct MRSP (Section 2.2).

2.1 Lower bound

We build our example on a 1-dimensional grid which can, of course, be considered as embedded in $\mathbb{R}^d$ for any $d \geqslant 1$. Let k be an integer, G_k denotes the set of integers $\{1, \ldots, k\}$ and $\mathcal{S}_k = G_k \cup (G_k + (k+1))$, that is, two copies of G_k separated by a gap of one point at $k+1$.

Proposition 1 *The set $\mathcal{S}_k$ has at least 2^{k-1} distinct MRSP.*

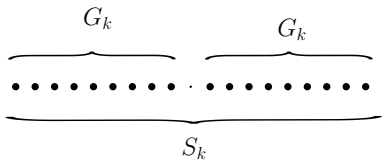

Figure 1: S_k is a set of $2k$ points on a 1-dimensional grid having at least 2^{k-1} distinct MRSP.

We show that any subset $P \subseteq G_k$ is a MRSP by arguing that for any $p^* \in \mathcal{S}_k \setminus P$, one of the translations that keeps P in $\mathcal{S}_k$ sends p^* either to $\{k+1\}$ or outside of $\mathcal{S}_k$. Indeed, let $Q \subseteq \mathcal{S}_k$ be a proper super-set of P and $p^* \in Q \setminus P$. If $p^* \geqslant k+2$ then $P + (k+1) \subseteq \mathcal{S}_k$ and $Q + (k+1) \not\subseteq \mathcal{S}_k$. If $p^* \leqslant k$ then $P + (k+1-p^*) \subseteq \mathcal{S}_k$ and $Q + (k+1-p^*) \not\subseteq \mathcal{S}_k$. This proves that any subset $P \subseteq G_k$ is a MRSP of $\mathcal{S}_k$. No translation can map a subset of G_k that contains 1 to another subset of G_k that contains 1, so all the subsets of G_k containing 1 are distinct. Therefore, at least 2^{k-1} of the subsets of S_k are distinct MRSP.

2.2 Upper bound

Let $\mathcal{S} = \{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ be a set of n points and $\mathcal{T} \subseteq \mathbb{R}^d$ the *set of translations* defined by

$$\mathcal{T} := \mathcal{S} - \mathcal{S} = \{x - y \mid (x, y) \in \mathcal{S}^2\}.$$

Both the points in $\mathcal{S}$ and the translations in $\mathcal{T}$ are *ordered lexicographically* as vectors of d real numbers, in the sense that if $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, then $x < y$ if $x_1 < y_1$ or $x_1 = y_1, \cdots, x_r = y_r$ and $x_{r+1} < y_{r+1}$ for some $r = 1, \ldots, d-1$. Let $\mathcal{A}$ denote the family of all *first* occurrences of subsets of $\mathcal{S}$ that are MRSP. By "first" we mean that a MRSP P is in $\mathcal{A}$ if and only if no translation $t < \mathbf{0}$ satisfies $P + t \subseteq \mathcal{S}$. We choose one representative of each equivalence class of MRSP under translation, so the number of distinct MRSP of $\mathcal{S}$ is $|\mathcal{A}|$. The following function maps each pattern to its set of translations:

$$\phi : \left\{ \begin{array}{rcl} 2^{\mathcal{S}} & \to & 2^{\mathcal{T}} \\ P & \mapsto & \{t \in \mathcal{T} \mid P + t \subseteq \mathcal{S}\} \end{array} \right.$$

For any repeated sub-pattern P, $|\phi(P)| \geq 2$ and if $P \in \mathcal{A}$ then $t \geq \mathbf{0}$ for every $t \in \phi(P)$.

For $1 \leqslant i \leqslant j \leqslant n$ let $\mathcal{A}_{ij} = \{P \in \mathcal{A} \mid \{a_i, a_j\} \subseteq P \subseteq \{a_i, \ldots, a_j\}\}$ be the set of all occurrences of MRSP spanning the range $\{a_i, \ldots, a_j\}$ and $\mathcal{T}_{ij} = \{t \in \mathcal{T} \mid t \geq \mathbf{0} \text{ and } \{a_i, a_j\} \subseteq \mathcal{S} \cap (\mathcal{S} - t)\}$ be the set of all non-negative translations compatible with a_i and a_j. Note that $\{\mathcal{A}_{ij}\}$ is a partition of $\mathcal{A}$, $\mathcal{A}_{11} = \{a_1\}$ and $\mathcal{A}_{ii}$ is empty for $i \geq 2$. So we have

$$|\mathcal{A}| = 1 + \sum_{1 \leqslant i < j \leqslant n} |\mathcal{A}_{ij}|. \tag{1}$$

We can now prove our upper bound.

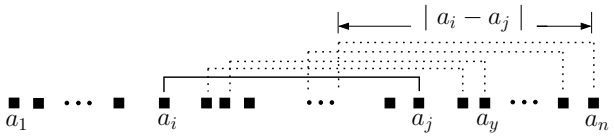

Figure 2: Bounding $|\mathcal{T}_{ij}|$ in 1-dimensional case; the same reasoning holds in $\mathbb{R}^d$ thanks to the total ordering.

Proposition 2 *A set of n points has at most $16 \cdot 2^{\lceil n/2 \rceil}$ distinct MRSP.*

Let P_1 and P_2 be two MRSP such that $\phi(P_1) = \phi(P_2)$. Then $\phi(P_1 \cup P_2) = \phi(P_1) = \phi(P_2)$ which leads to $P_1 \cup P_2 = P_1$, since P_1 is a MRSP, and $P_1 \cup P_2 = P_2$, as P_2 is also a MRSP. Thus, ϕ defines an injection from $\mathcal{A}$ on the subsets of $\mathcal{T}$. If $P \in \mathcal{A}_{ij}$ then $\phi(P) \subseteq \mathcal{T}_{ij}$ and ϕ induces an injection from $\mathcal{A}_{ij}$ on the subsets of $\mathcal{T}_{ij}$. Hence,

$$|\mathcal{A}_{ij}| \leqslant 2^{|\mathcal{T}_{ij}|}.$$

For each $t \in \mathcal{T}_{ij} \setminus \{\mathbf{0}\}$, $t > \mathbf{0}$ and there exists unique $y > j$ such that $a_j + t = a_y$. Hence, $|\mathcal{T}_{ij} \setminus \{\mathbf{0}\}| \leqslant n - j$. Because $\phi(P \in \mathcal{A}_{ij})$ includes $\mathbf{0}$ and at least one translation $t \in \mathcal{T}_{ij} \setminus \{\mathbf{0}\}$, it follows that

$$|\mathcal{A}_{ij}| \leqslant 2^{n-j} - 1.$$

As any MRSP in $\mathcal{A}_{ij}$ corresponds to a subset of $\{a_{i+1}, \ldots, a_{j-1}\}$ we also have that

$$|\mathcal{A}_{ij}| \leqslant 2^{j-i-1}.$$

Applying these to equation (1), we get

$$|\mathcal{A}| \leqslant 1 + \sum_{1 \leqslant i < j \leqslant n} 2^{\min(n-j, j-i-1)}.$$

Splitting the sum at $j = \lceil \frac{n+i}{2} \rceil + 1$, we have

$$|\mathcal{A}| \leqslant 1 + 2\sum_{i=1}^{n} \sum_{j=i+1}^{\lceil \frac{n+i}{2} \rceil + 1} 2^{j-i-1} \leqslant 1 + 2\sum_{i=1}^{n} 2^{\lceil \frac{n-i}{2} \rceil + 1} \leqslant 1 + 8\sum_{\ell=1}^{\lceil \frac{n}{2} \rceil} 2^{\ell}$$

and finally $|\mathcal{A}| \leqslant 16 \cdot 2^{\lceil n/2 \rceil}$.

References

[1] P. Brass. Combinatorial geometry problems in pattern recognition. *Discrete and Computational Geometry*, 28:495–510, 2002.

[2] P. Brass, W. Moser, and J. Pach. *Research Problems in Discrete Geometry*. Springer-Verlag, 2005.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399