Approximation Guarantees for Max Sum and Max Min Facility Dispersion with Parameterised Triangle Inequality and Applications in Result Diversification

Marcin Sydow[+,*]

[+] Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[*] Web Mining Lab, Polish-Japanese Institute of Information Technology, Warsaw, Poland
msyd@ipipan.waw.pl

Abstract. Facility Dispersion Problem, originally studied in Operations Research, has recently found important new applications in Result Diversification approach in information sciences. This optimisation problem consists in selecting a small set of p items out of a large set of candidates to maximise a given objective function. The function expresses the notion of *dispersion* of a set of selected items in terms of a pair-wise *distance* measure between items.
In most known formulations the problem is NP-hard, but there exist 2-approximation algorithms for some cases if distance satisfies triangle inequality.
We present generalised $2/\alpha$ approximation guarantees for the Facility Dispersion Problem in its two most common variants: Max Sum and Max Min, when the underlying dissimilarity measure satisfies *parameterised triangle inequality* with parameter α. The results apply to both relaxed and stronger variants of the triangle inequality.
We also demonstrate potential applications of our findings in the result diversification problem including web search or entity summarisation in semantic knowledge graphs, as well as in practical computations on finite data sets.

Keywords: diversity, max sum and max min facility dispersion, approximation algorithms, parameterised triangle inequality

1 Introduction

The concept of *diversity-awareness* has important practical applications in web search, recommendation, database querying or summarisation (e.g. [11, 5, 17]). The general idea is to return to the user the set of items (being database query or search results, recommended items, etc.) that are not only *relevant* to the query but also *diversified*. The rationale behind such approach is to reduce the risk of *result redundance* and to cover as many different aspects of the query as possible. Equivalently, it is a technique for maximising the likelihood that the user's *unkown intent* behind a potentially *ambiguous query* is satisfied, at least partially.

One possible formulation of the described *diversification problem* is by treating it as the *Facility Dispersion* optimisation problem that was originally studied in Operations Research (e.g. [15, 6, 13, 7]). More precisely, the problem concerned selecting locations for some dangerous or obnoxious facilities in order to make them *mutually distant* to each other. The range of possible applications is wide, and includes minimising the effects of a terroristic or military attack (if items represent nuclear plants, amunition dumps, etc.) or to avoid self-competition between the stores of the same brand, etc.

In the context of information sciences, the notion of mutual spatial distance has been substituted with that of *pairwise dissimilarity* between the items to be returned to the user.

Examples of recent applications of this approach include the *Result Diversification Problem* in web search (e.g. [5, 11]) or in diversity-aware graphical summarisation of entities in semantic knowledge graphs [17, 14], for example.

A quite recent interesting work [5] mentions in its very last sentence the issue of studying the relationship between the approximation ratio of the Max Sum Diversification Problem and the parameter of the relaxed triangle inequality among the interesting future research problems but does not address it.

Therefore, our results presented in this article might be of additional interest in this context.

The facility dispersion problem is NP-hard in most common variants, in particular, in Max Min and Max Sum variants studied in this article, and it remains such even when the distance (dissimilarity) function d satisfies the triangle inequality [8, 18]. However, in such a case, there are polynomial-time approximation algorithms for this problem with approximation factor of 2 [13, 15].

This article focuses on two most common variants of the problem called *Max Sum* and *Max Min Facility Dispersion*. The main results include the generalisations of approximation guarantees for metric cases [13, 15] to the factor of $2/\alpha$ for the case when the distance function satisfies *parameterised triangle inequality* with parameter α.

1.1 Contributions

This article is a substantial extension of the short paper [16], where the following results were first mentioned:

- generalised approximation guarantee for Max Sum Dispersion Problem with Parameterised Triangle Inequality
- the link between the above result and the *Result Diversification Problem* (e.g. [11, 5]) that is of interest in web search and other applications
- observation that parameterised triangle inequality is satisfied by finite datasets

Compared to the short paper [16] this article presents the results concerning generalised approximation guarantee for Max Sum Dispersion Problem with Parameterised Triangle Inequality and other results mentioned above, in its full, extended form with a deeper and wider disscussion and, in addition, presents some additional, completely novel results concerning another commonly studied variant of the problem, known as Max Min Facility Dispersion. It also contains the proofs of the presented theorems. The list of the novel contributions includes:

- discussion of the related work concerning Max Sum, Max Min and related problems (Section 2)
- some remarks on the parameterised triangle inequality (Section 4)

- tight example for the approximation guarantee for Max Sum Dispersion Problem generalised for parameterised triangle inequality (Section 5.2)
- generalised approximation guarantee for Max Min Dispersion Problem with parameterised triangle inequality (Section 5.3)
- generalised impossibility result (i.e. lower-bound approximation factor) for the Max Min Dispersion Problem with parameterised triangle inequality (Section 5.4) with proof (Appendix)

2 Related Work

Up to the best knowledge of the author, no results for the impact of any form of parameterising the triangle inequality on the approximation guarantees of the Max Sum or Max Min Dispersion Problem have been published at the time of writing this (except our short paper [16], mentioned before).

2.1 Related Work on Result Diversification

Result diversification approach has recently become an intensively studied topic in web search, recommendation, databases or summarisation.

The connection between the Facility Dispersion Problem and the Result Diversification approach was presented in [11] by proposing an appropriate transformation of the dissimilarity function. In this paper, in Section 6.2, we observe that our result holds when this transformation is applied.

In particular, Result Diversification based on Max Sum was recently proposed in [14] in a novel context of diversity-aware entity summarisation in semantic knowledge graphs [17].

The results contained in this paper are related to the question recently stated in the last sentence of [5], where the Max Sum Diversification Problem was studied in a more general framework of the monotone submodular functions under matroid constraints. The question concerned the impact of relaxing triangle inequality for approximation guarantee for a generalised Max Sum Dispersion Problem. Our result can be viewed as a partial answer to the special case of this problem.

2.2 Previous Work on Facility Dispersion Problem

Facility Dispersion Problem was studied in many works in Operations Research.

In [15] there are presented approximation results for the metric Max Min (factor of 2 and its tightness, among others) and Max Average Facility Dispersion (factor of 4). Some of these results were improved (factor of 2 for metric k-dispersion) in [13] for the metric Max Sum variant. To be precise, the variant of the Max Sum Facility Dispersion Problem considered in this article is equivalent to the k-dispersion problem considered in [13] for the value of parameter $k = 1$. The meaning of the parameter k is explained later in the Section 3.1 (the only value of k that is relevant to our applications in web search diversification and recommendation, etc. is $k = 1$).

Many more variants of the Facility Dispersion problem (over 10) and approximation algorithms for them were studied in [6] and its extension [7]. It would be interesting to analyse the applicability of these variants in the result diversification problem in information sciences, however we are not aware of such work at the time of writing.

None of the works mentioned above considered parameterised variant of the distance in the problem specification. The most relevant results to ours in the above works concern 2-approximation results for Max Sum and Max Min Facility Dispersion Problem with metric distance.

Our main results presented here build on the results in the two works [15, 13] mentioned above, and the proofs presented in this paper are extensions of the proofs presented in these works and in [12] by appropriately introducing the parameter of the parameterised triangle inequality.

2.3 Previous Work on the Impact of Parameterising Triangle Inequality on Approximation Guarantees for Graph Optimisation Problems

The impact of various forms of parameterising the triangle inequality on the approximation guarantees for various hard optimisation problems, especially for Travelling Salesman Problem was studied previously in many works.

Most of such works consider parameterising the triangle inequality in the form of the so-called *ρ-relaxed triangle inequality* (Section 4.1). The both formulations are almost equivalent, the similarities and differences between the ρ-relaxed and the Parameterised Triangle Inequality formulation used in this paper are discussed in Section 4.1

One of the early works on the impact of relaxing triangle inequality on approximating the solution of the Travelling Salesman Problem (TSP) is [1]. It demonstrates an interesting phenomenon. Namely, some known approximation algorithms for TSP can benefit from the relaxed triangle inequality and give better approximation guarantees (e.g. quadratic improvement of the approximation guarantee with the relaxation parameter for the double-spanning tree heuristic) while other algorithms (e.g. Christofides algorithm) cannot benefit from it.

Improving the approximation quality of the TSP in the context of relaxed triangle inequality is also studied e.g. in [2], which presents, among others, 4τ approximation for relaxed triangle inequality with parameter τ. In [4] some improvements on approximation guarantees for TSP with sharpened (stronger) triangle inequality are presented. In [3] further improvements of approximation guarantees for TSP with strenghtened triangle inequlity for some specific range of values of the parameter are presented.

Other aspects related to the influence of relaxing the triangle inequality on the performance of algorithms are studied. For example, a recent work [10] concerns the impact of triangle inequality violations (resulting, for example, from rounding errors or network delays, etc.) on the quality of solution in the vehicle routing problem.

Many of these works consider either relaxed or sharpened variant of the triangle inequality. In this paper, we consider the full possible range of the parameter values.

An example of a practically useful pair-wise dissimilarity measure that naturally satisfies parameterised triangle inequality is a variant of the *NEM* (Non-linear Elastic Matching) measure used in pattern matching for comparing visual shapes [9].

3 Facility Dispersion Problem

In the Facility Dispersion Problem, the input consists of a complete, undirected graph $G(V, E)$, an edge-weight function $d : E \to R^+$ that represents *pairwise distance* between the vertices and a positive natural number $p \leq |V|$. The task is to select a p-element subset $P \subseteq V$ that maximises the objective function $f_d(P)$ that represents the notion of *dispersion* of the elements of the selected set P. In the remaining part of the article we will simplify the notation and use $f(P)$ instead of $f_d(P)$, since the pairwise distance function d will be known from the context.

Depending on the particular form of the objective dispersion function $f(P)$ to be maximised there are considered several variants of the Facility Dispersion Problem. The two most commonly studied variants are Max Sum and Max Min Facility Dispersion and are described in Sections 3.1 and 3.2, respectively.

3.1 Max Sum, or equivalently, Max Average Facility Dispersion Problem

In the Max Sum Dispersion Problem the objective function to be maximised is defined as follows:

$$f_{SUM}(P) = \sum_{\{u,v\} \subseteq P} d(u, v). \tag{1}$$

Relation to k-dispersion Problem. The above formulation of the problem is equivalent to the *k-dispersion* problem that was studied, for example, in [13] for $k = 1$. In the k-dispersion problem, for $k \in \{1, ..., \lfloor |V|/p \rfloor\}$ there are k different p-element sets $P_1, ..., P_k$, $|P_i| = p$ to be found that maximise the following objective function:

$$\sum_{i=1}^{k} \sum_{\{u,v\} \subseteq P_i} d(u, v) \tag{2}$$

Since our main motivation is in the application of the problem in the *Result Diversification* problem (as explained in Section 6) and other related problems (e.g. recommendation or summarisation problems) where there is exactly *one* set of elements to be diversified, we focus only on the case $k = 1$ in this work.

Relation to Average Sum Dispersion Problem. In some works (e.g. [15, 12]) the objective function is formulated as:

$$f_{AVE}(P) = \frac{2}{p(p-1)} \sum_{\{u,v\} \subseteq P} d(u, v),$$

and then the problem is known as *Max Average* Facility Dispersion. Since, for the fixed value of p, the formulation is obviously *equivalent* to that of Max Sum Facility Dispersion, we will refer to them interchangeably in this paper.

INPUT: An undirected graph $G(V, E)$ with edge-weight function $d : E \to R^+$, a natural number $1 < p \leq |V|$

OUTPUT: A p-element set $P \subseteq V$

1. $P = \emptyset$
2. Compute a maximum-weight $\lfloor p/2 \rfloor$-matching M^* in G
3. For each edge in M^*, add its both ends to P
4. In case p is odd, add any node from $V \setminus P$ to P
5. return P

Fig. 1. HRT: an efficient 2-Approximation Algorithm for Max Average Dispersion Problem

The Max Sum Dispersion problem is NP-hard even if the distance function d is a metric, but in such case there exists a polynomial-time algorithm of approximation factor of 2 that was presented in [13].

Figure 3.1 shows a heuristic, polynomial-time approximation algorithm, based on computing a maximum-weight matching, that guarantees approximation factor of 2 for Max Average Dispersion when d satisfies triangle inequality. It was presented in [13] under the name HRT. A straightforward implementation of the algorithm makes its time complexity $O(|V|^3)$, however it is possible to implement it so that the time complexity is reduced to $O(|V|^2(p + log(|V|))$. There exists another algorithm for the same problem, for which the same approximation guarantee of 2 can be proved, but which has better time complexity of $O(|V|^2 + p^2 log(p))$. For all the remaining details we refer the reader to [13].

3.2 Max Min Dispersion Problem

In the Max Min Dispersion Problem the objective function to be maximised is defined as follows:

$$f_{MIN}(P) = min_{u,v \in P, u \neq v} d(u, v). \tag{3}$$

Similarly to Max Sum, even if the distance function d satisfies the standard triangle inequality, the problem is NP-hard, but in such case there exists a polynomial-time algorithm that guarantees an approximation factor of 2.

The algorithm and its 2-approximation guarantee proof is presented in [15] under the name *GMM* and is shown in Figure 2. Its time complexity is $O(p^2|V|)$.

4 Parameterised Triangle Inequality

Assume that V is a non-empty universal set and $d : V^2 \to R^+ \cup \{0\}$ is a distance function. More preciselty, it is assumed that d for all $u, v \in V$ satisfies the properties of *discernibility* ($d(u, v) = 0 \Leftrightarrow u = v$) and *symmetry* ($d(u, v) = d(v, u)$). If, in addition, INPUT: An undirected graph $G(V, E)$ with edge-weight function $d : E \to R^+$, a natural number $1 < p \leq |V|$

OUTPUT: A p-element set $P \subseteq V$

1. $P = argmax_{\{v_i,v_j\} \subseteq V} d(v_i, v_j)$
2. while ($|P| < p$):
   - find $v \in V \setminus P$ so that $v = argmax_{v \in V \setminus P}(min_{u \in P}\{d(v, u)\})$
   - $P = P \cup \{v\}$
3. return P

Fig. 2. GMM: an efficient 2-Approximation Algorithm for Max Min Dispersion Problem

for all mutually different $u, v, z \in V$, d satisfies the *triangle inequality*: $d(u, v) + d(v, z) \geq d(u, z)$, d is called a *metric*.

Here, we introduce the following definition of *parameterised triangle inequality*:

Definition 1. *Let V be a set, $\alpha \in R$, $0 \leq \alpha \leq 2$. A distance function $d : V^2 \to R^+ \cup \{0\}$ satisfies* parameterised triangle inequality *(α-PTI) with parameter α iff for all mutually different $u, v, z \in U$:*

$$d(u, v) + d(v, z) \geq \alpha d(u, z).$$

α-PTI generalises the standard triangle inequality (for $\alpha = 1$). 0-PTI is the weakest variant, i.e. it is satisfied by any distance function d, which is equivalent to the *semi-metric*. Observe that the higher the value of the α parameter, the stronger the property.

The value of α cannot be higher than 2:

Lemma 1. *The value of 2 is the highest possible value for α in α-PTI.*

Proof. Assume that some distance function $d : V^2 \to R^+ \cup 0$ satisfies α-PTI for $\alpha > 2$ with $|V| \geq 3$. Let u, v, z be three different elements of V. Let introduce the following denotations: $a = d(u, v), b = d(v, z), c = d(z, u)$. We obtain directly from the definition that $a + b > 2c$, and hence $c < \frac{a+b}{2}$.

By summing up the following two inequalities obtained directly from the definition: $b+c > 2a$ and $c+a > 2b$ and then using the above observation, we obtain the following chain of inequalities (that makes a contradiction):

$$2a + 2b < 2c + a + b < 2\frac{a + b}{2} + a + b = 2a + 2b.$$

(Quod erat demonstrandum)$\diamondsuit$

Notice that 2-PTI implies that all non-zero distances are equal, that is equivalent to being a *discrete metric* (up to rescaling), hence the Facility Dispersion Problem becomes trivial for the case $\alpha = 2$.

4.1 PTI vs Relaxed Triangle Inequality

In some works (e.g. [9] or the works mentioned in Section 2.3) the concept of ρ-relaxed triangle inequality (RTI) is considered instead of *parameterised triangle inequality*. It is defined as follows, for some $\rho \in R^+$:

$$\rho(d(u,v) + d(v,z)) \geq d(u,z).$$

Obviously, such formulation is equivelent to α-PTI with $\rho = 1/\alpha$ with the following observations:

- the case of semi-metric ($\alpha = 0$) cannot be expressed by ρ-relaxed triangle inequality
- the range of possible values of α is bounded $[0, 2]$, with the metric case being exactly in the middle of that range ($\alpha = 1$). For RTI its parameter value has no upper bound and cannot be smaller than $\frac{1}{2}$.
- for PTI the higher the value of the parameter, the stronger the property, while for the RTI it is the opposite.

For the above reasons the α-PTI formulation seems a bit more natural than that of RTI.

5 Improved Approximation Guarantees for α-PTI Case

In this section we present the main results, i.e. we demonstrate that the approximation guarantees for Max Sum and Max Min Facility Dispersion problems can be generalised from the value of 2 (for the metric case) to the value of $2/\alpha$ when the distance function d satisfies α-PTI for $0 \leq \alpha \leq 2$. The longer proofs are presented in the Appendix.

5.1 $2/\alpha$ Approximation Guarantee for Max Sum Dispersion Satisfying α-PTI

In this subsection we present a theorem that generalises and extends the 2-factor approximation guarantee of the algorithm presented in Figure 3.1 for the case when the function d satisfies α-PTI. The algorithm, under the name HRT, was presented in [13] together with a proof for a metric case. Our proof of this theorem, is an adaptation of the one presented in [12] by properly introducing the α parameter and is presented in the Appendix.

Theorem 1. *Let I be an instance of Max Average Dispersion problem with distance function d satisfying α-PTI for $0 < \alpha < 2$. Let $OPT(I)$ denote the value of an optimal solution and by $HRT(I)$ the value of the solution found by the algorithm HRT. It holds that $OPT(I)/HRT(I) < 2/\alpha$.*

As mentioned in Section 3.1, the result applies equivalently to the Max Sum Dispersion problem. For the value of $\alpha = 2$ the problem becomes trivial, since d becomes a discrete metric in such case.

5.2 Tight Example for Max Sum Dispersion

We show here that the $2/\alpha$ bound for the algorithm for Max Sum Dispersion with α-PTI presented in Figure 3.1 is (asymptotically) tight.

Let the graph $G(V, E)$, where $|V| \geq 2p$, $M = 2/\alpha$, $m = \beta M = 2\beta/\alpha$, for some $0 < \beta < 1$ (intentionally, close to 1)contain exactly $\lfloor p/2 \rfloor$ edges of weight M and exactly one p-clique, call it C, within which all the edges have weights of m, and all the other edges have weight of 1. Note that α-PTI is satisfied.

The HRT algorithm will select the ends of the $\lfloor p/2 \rfloor$ edges of weight M and, in case p is odd, any arbitrary vertex that brings a total weight of $(p - 1)$.

But the optimum, for β being sufficiently close to 1, is the set of vertices of the p-clique C, with the value of $OPT = p(p-1)m/2$. Thus, we have:

$$OPT/H_{even} = \frac{\frac{p(p-1)}{2}m}{\frac{p(p-1)}{2} + \frac{p}{2}(M-1)} = \frac{m}{1 + \frac{(M-1)}{(p-1)}}, \quad (4)$$

$$OPT/H_{odd} = \frac{\frac{p(p-1)}{2}m}{\frac{(p-1)(p-2)}{2} + \frac{(p-1)}{2}(M-1) + (p-1)} = \frac{m}{1 + \frac{(M-1)}{p}}. \quad (5)$$

Both the above expressions, for sufficiently large p are arbitrarily close to $m = 2\beta/\alpha$ that is arbitrarily close to $2/\alpha$ for $\beta < 1$ sufficiently close to 1.

5.3 $2/\alpha$ Approximation Guarantee for Max Min Dispersion Satisfying α-PTI

The following theorem is a generalisation and extension of the result [15] of 2-approximation for Max Min Dispersion satisfying standard triangle inequality.

Theorem 2. *Let I be an instance of the Max Min Dispersion Problem where the distance function d satisfies α-PTI for $0 < \alpha \leq 2$. Let OPT(I) denote the optimum value of the objective function f_{MIN} (Equation 3) for this instance and GMM(I) denote the value of the solution found by the GMM algorithm for I (Figure 2). Then $OPT(I)/GMM(I) \leq 2/\alpha$, i.e. the GMM algorithm provides $2/\alpha$-approximation guarantee for this problem.*

The proof is presented in Appendix 8.2 and constitutes our extension of the one presented in [15] by properly introducing the parameter α.

5.4 Lower Bound for Max Min Dispersion with α-PTI

For metric cases, the 2-factor polynomial approximation algorithm is the best that exists for Max Min Dispersion problem, assuming $P \neq NP$ [15].

Below, we present a generalisation of this fact to the case of α-PTI.

Theorem 3. *Assume that the distance function d satisfies α-PTI for $0 < \alpha < 2$, and that $\beta < 2/\alpha$ is a real positive number. There is no poly-time algorithm for Max Min Dispersion with approximation factor of β unless $P = NP$.*

Proof. Imagine an instance $(G(V, E), p)$ of the Maximum Independent Set problem, stated in its decision version: "Does there exist an independent set of size at least p in the given graph G"? Let us set weights of edges in E to $2/\alpha$ and add all the other possible edges of weight 1. A polynomial-time β-approximation algorithm for Max Min Dispersion problem on such modified graph would return the value of 1 iff the answer for the question is positive i.e. solve an NP-hard problem. *(Quod erat demonstrandum)* ♢

One can notice that for the value of $\alpha = 2$, as explained in Section 4, all the pairwise distances in V are the same, and all feasible solutions to Max Min Dispersion have the same value. Thus, formally, the guarantee of approximation factor of $2/\alpha = 1$ also holds.

6 Practical Applications

6.1 α-PTI in Practical Computational Problems

In practical applications, the input dataset V is always finite. Here, we make the observation that this fact implies that the distance function d *always* satisfies α-PTI for some $0 < \alpha \leq 2$. To practically find the *actual maximum value* of the α parameter it suffices to check all triples $u, v, z \in V$ in $O(|V|^3)$ time, so that the α-PTI is satisfied as follows: $\alpha = min_{u,v,z \in V, |\{u,v,z\}|=3}[d(u, v) + d(v, z)]/d(u, z)$.[1]

In this way, it is possible to guarantee constant-factor approximation algorithms for Max Sum and Max Min Dispersion problems even if no other theoretical properties about the distance function are known. In particular, in the metric case, it is *always* possible to guarantee better than 2 approximation factor for finite data sets unless there are no "co-linear" points in the data.[2]

In the *on-line* variant of the considered problems (i.e. when data comes in time), α can be systematically updated while data comes.

6.2 Applications to the Result Diversification Problem in Web Search, etc.

In this section, it is demonstrated how the theoretical results from section 5.1 impact some important recent applications in information sciences including web search and others.

In web search, the problem known as *Result Diversification Problem* can be specified as follows [11]. There is given a set V of documents that are potentially relevant to a user query, a number $p \in N^+, p < |V|$, a *document relevance* function $w : V \to R^+$ and a *document pairwise dissimilarity* function $d : V^2 \to R^+ \cup \{0\}$. The task in this problem is to select a subset $P \subseteq V$ that maximises the properly defined *diversity-aware set relevance function*. For example, in [11] the following objective function (to be maximised) is proposed as the diversity-aware relevance function:

[1] Notice that u, v, z are pairwise different and $d(u, v)$ is always positive for $u \neq v$

[2] I.e. pair-wise different u, v, z so that $d(u, z)$ is exactly equal to $d(u, v) + d(v, z)$

$$f_{div-sum}(\lambda, P) = (p-1)\sum_{v \in P} w(v) + 2\lambda \sum_{\{u,v\} \subseteq P} d(u,v). \quad (6)$$

The $\lambda \in R^+ \cup \{0\}$ parameter controls the balance between the relevance term and the diversity term.

It is explained in the same work that by a proper modification of d to d' (Equation 7) it is possible to make the described problem of maximising $f_{div-sum}(\lambda, P)$ equivalent to maximising $\sum_{\{u,v\} \subseteq P} d'_\lambda(u,v)$, where:

$$d'_\lambda(u,v) = w(u) + w(v) + 2\lambda d(u,v). \quad (7)$$

In this way, the *result diversification* problem described above is equivalent to the Max Sum Dispersion problem for d'_λ. [3]

It is also claimed in [11] that d' is a metric if d is such. [4]

The following lemma 2 extends the application of the results presented in this paper to the *result diversification* problem on any finite datasets, due to the observation in Section 6.1.

Lemma 2. *If distance function* d *satisfies* α*-PTI, for some* $0 < \alpha \leq 1$*, then the modified distance function* d'_λ *defined in equation 7 also satisfies* α*-PTI.*

Proof. $d'_\lambda(u,v) + d'_\lambda(v,z) = 2w(v) + w(u) + w(z) + 2\lambda(d(u,v) + d(v,z)) \geq$

$$\geq 2w(v) + w(u) + w(z) + 2\lambda\alpha d(u,z) \geq$$

$\geq 2w(v) + \alpha[w(u)/\alpha + w(z)/\alpha + 2\lambda d(u,z)] \geq \alpha d'_\lambda(u,z)$. *(Quod erat demonstrandum)* ◇

The diversification problem has other interesting applicaitons than in web search. For example, the problem of optimising the $f_{div-sum}(\lambda, P)$ objective function has been recently adapted to the novel context of *diversified entity summarisation in knowledge graphs* [17, 14].

---

[3] Considering the Max Min Dispersion Problem, [11] also considers a variant of bi-criteria objective function: $f_{div-min}(\lambda, P) = min_{u \in P} w(u) + \lambda min_{u,v \in P} d(u,v)$ and the authors seem to suggest that maximising it is equivalent to the Max Min Dispersion by defining a modified distance function $d'_\lambda(u,v) = (w(u)+w(v))/2 + \lambda d(u,v)$. But this is, unfortunately, not true, in general. To see this consider the following example: $V = \{u,v,z\}, w(u) = 1, w(v) = 1, w(z) = 20, d(u,v) = 10, d(v,z) = d(z,u) = 1, p = 2, \lambda = 1$. In this example, $f_{div-min}(\lambda, P)$ is maximised for $P = \{u,v\}$ but $f'_{div-min}(\lambda, P) = min_{u \neq v \in P} d'_\lambda(u,v)$ is maximised for $P = \{u,z\}$ or $P = \{v,z\}$. Furthermore, it is necessary to make the discernibility property satisfied by explicitly setting $d'_\lambda(u,v) = 0$ for $u = v$ since this condition is necessary in the proof of approximation factor of 2 that authors of [11] refer to (see the proof of Theorem 2, in the Appendix, more precisely, the property of non-emptiness of S_i^*).

[4] Formally, this is not true in the form proposed in [11] because the discernibility property would be not satisfied by d' (see Equation 7). A simple explicit addition of $d'(u,v) = 0$ for $u = v$, however, makes this claim correct.

7 Conclusions and Future Work

In this work we gathered the results concerning approximation guarantees for metric cases of Max Min and Max Sum Dispersion problems and generalised them to the case of parameterised triangle inequality.

It was demonstrated that they may have additional practical impact on computations on real, finite datasets and, in particular, on recent important diversity-related applications in information sciences as in [11] or [14], for example.

In addition, one can notice that our results concerning the impact of parameterising the triangle inequality on the approximation guarantees in the Max Sum Dispersion problem answer the open research question mentioned quite recently in the last sentence of [5] for the special case of the uniform matroid and simple form of the objective function. It would be interesting to extend this answer for more general cases.

The way in which α-PTI property affects the guarantees of approximation algorithms and their proofs presented in this paper is quite simple, i.e. the value of approximation factor of 2 changes to the value of $2/\alpha$. It is important to notice, however, that there exist natural optimisation problems, in which generalising the triangle inequality to α-PTI does not affect the approximation factor in such a straightforward way. A good example is the *Travelling Salesman Problem* in which one approximation algorithm (double-spanning tree heuristic) can be modified so that the approximation guarantee can benefit from relaxing the triangle inequality while the Christofides algorithm for TSP cannot be mofified in the same way [1].

In the context of practical computations, it would be interesting to study which distance metrics used in practical applications (such as the measure described in [9]) satisfy α-PTI and what is the value of the parameter.

Acknowledgments

The author would like to thank the anonymous reviewers, Professor J.Koronacki (the editor) and Professor M.A.Kłopotek from the Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland, for their valuable comments on the manuscript that made it possible to improve this article.

Work supported by Polish National Science Centre grant "DISQUSS" 2012/07/B/ST6/01239

References

1. Thomas Andreae and Hans-Jurgen Bandelt. Performance guarantees for approximation algorithms depending on parametrized triangle inequalities. *SIAM Journal of Discrete Mathematics*, 8:1–16, 1995.
2. Michael A. Bender and Chandra Chekuri. Performance guarantees for the TSP with a parameterized triangle inequality. *Information Processing Letters*, 73(12):17 – 21, 2000.
3. Markus Blser. An improved approximation algorithm for the asymmetric tsp with strengthened triangle inequality. In JosC.M. Baeten, JanKarel Lenstra, Joachim Parrow, and GerhardJ. Woeginger, editors, *Automata, Languages and Programming*, volume 2719 of *Lecture Notes in Computer Science*, pages 157–163. Springer Berlin Heidelberg, 2003.

4. Hans-Joachim Bockenhauer, Juraj Hromkovic, Ralf Klasing, Sebastian Seibert, and Walter Unger. Approximation algorithms for the TSP with sharpened triangle inequality. *Information Processing Letters*, 75(3):133 – 138, 2000.
5. Allan Borodin, Hyun Chul Lee, and Yuli Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In *Proceedings of the 31st symposium on Principles of Database Systems*, PODS '12, pages 155–166, New York, NY, USA, 2012. ACM.
6. Barun Chandra and Magnus M. Halldorsson. Facility dispersion and remote subgraphs. In Rolf Karlsson and Andrzej Lingas, editors, *Algorithm Theory SWAT'96*, volume 1097 of *Lecture Notes in Computer Science*, pages 53–65. Springer Berlin Heidelberg, 1996.
7. Barun Chandra and Magnus M Halldorsson. Approximation algorithms for dispersion problems. *Journal of Algorithms*, 38(2):438 – 465, 2001.
8. Erhan Erkut. The discrete p-dispersion problem. *European Journal of Operational Research*, 46(1):48 – 60, 1990.
9. Ronald Fagin and Larry Stockmeyer. Relaxing the triangle inequality in pattern matching. *Int. J. Comput. Vision*, 30(3):219–231, December 1998.
10. Christopher L. Fleming, Stanley E. Griffis, and John E. Bell. The effects of triangle inequality on the vehicle routing problem. *European Journal of Operational Research*, 224(1):1 – 7, 2013.
11. Sreenivas Gollapudi and Aneesh Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 381–390, New York, NY, USA, 2009. ACM.
12. Teofilo F Gonzalez. *Handbook of approx. algorithms and metaheuristics*. CRC Press, 2007.
13. Refael Hassin, Shlomi Rubinstein, and Arie Tamir. Approximation algorithms for maximum dispersion. *Oper. Res. Lett.*, 21(3):133–137, 1997.
14. Witold Kosiński, Tomasz Kuśmierczyk, Pawel Rembelski, and Marcin Sydow. Application of ant-colony optimisation to compute diversified entity summarisation on semantic knowledge graphs. In *Proc. of International IEEE AAIA 2013/FedCSIS Conference, Annals of Computer Science and Information Systems*, volume 1, pages 69–76, 2013.
15. S.S.Ravi, D.J.Rosenkrantz, and G.K.Tayi. Heuristic and special case algorithms for dispersion problems. *Operations Research*, 42(2):299–310, 1994.
16. Marcin Sydow. Improved approximation guarantee for max sum diversification with parameterised triangle inequality. In Troels Andreasen, Henning Christiansen, Juan-Carlos Cubero, and Zbigniew Ras, editors, *Foundations of Intelligent Systems*, volume 8502 of *Lecture Notes in Computer Science*, pages 554–559. Springer International Publishing, 2014.
17. Marcin Sydow, Mariusz Pikula, and Ralf Schenkel. The notion of diversity in graphical entity summarisation on semantic knowledge graphs. *Journal of Intelligent Information Systems*, 41:109–149, 2013.
18. D. W. Wang and Yue-Sun Kuo. A study on two geometric location problems. *Inf. Process. Lett.*, 28(6):281–286, August 1988.

8 Appendix: Proofs

In this Appendix we present two longer proofs of theorems presented in previous sections.

8.1 Proof of Theorem 1 from Section 5.1

The following, including the Lemmas, Theorem and their proofs constitute our extensions of the versions presented in [12][pp. 38-8–38-9] (earlier variants were in [13]). The extensions presented here consist mostly in properly introducing the parameter α.

Let us introduce some denotations. Let $V' \subseteq V$ be a non-empty subset of vertices. Let $G(V')$ denote the complete graph induced on V' and $W(V')$,$W'(V')$ denote the total weight and average weight of edges in $G(V')$ respectively. By analogy, for a non-empty subset $E' \subseteq E$ of edges, let denote by $W(E')$ and $W'(E') = W(E')/|E'|$ the total and average weight of the edges in E', respectively. We use the following technical Lemma, presented in [13].

Lemma 3. *If $V' \subseteq V$ is a subset of vertices of cardinality at least $p \geq 2$ and M'^{*} is a maximum-weight $\lfloor p/2 \rfloor$-matching in $G(V')$, then $W'(V') \leq W'(M'^{*})$.*

A very short proof, presented in [13] does not assume *anything* on the distance function d, so that we omit it here.

Lemma 4. *Assume that the distance function d satisfies α-PTI for some $0 < \alpha < 2$. If $V' \subseteq V$ is a subset of $p \geq 2$ vertices and M is any $\lfloor p/2 \rfloor$-matching in $G(V')$, then $W'(V') > (\alpha/2)W'(M)$.*

Proof. (of Lemma 4) Let $M = \{\{a_i, b_i\} : 1 \leq i \leq \lfloor p/2 \rfloor\}$ and let denote by V_M the set of all vertices that are ends of the edges in M. For each edge $\{a_i, b_i\} \in M$ let E_i denote the set of edges in $G(V')$ that are incident on a_i or b_i, except the edge $\{a_i, b_i\}$ itself. From α-PTI we get that for any vertex $v \in V_M \setminus \{a_i, b_i\}$ we have $d(v, a_i) + d(v, b_i) \geq \alpha d(a_i, b_i)$. After summing this inequality over all the vertices in $V_M \setminus \{a_i, b_i\}$ we obtain:

$$W(E_i) \geq \alpha(p-2)d(a_i, b_i). \tag{8}$$

There are two cases:

Case 1: p is even, i.e. $\lfloor p/2 \rfloor = p/2$. After summing up the Inequality 8 above, over all the edge sets E_i, $1 \leq i \leq p/2$, we obtain, on the left-hand side, each edge of $G(V')$ twice, except those in M. Thus, $2[W(V') - W(M)] \geq \alpha(p-2)W(M)$. If we substitute in the last inequality $W(V') = W'(V')p(p-1)/2$ and $W(M) = W'(M)p/2$, and divide both sides by p, we can quickly get to $W'(V') \geq (\alpha/2)W'(M)[p-2+(2/\alpha)]/(p-1)$, that is equivalent to $W'(V') > (\alpha/2)W'(M)$ (for $\alpha < 2$). This completes the proof for the Case 1.

Case 2: p is odd, i.e. $\lfloor p/2 \rfloor = (p-1)/2$. Let x be the only node in $V' \setminus V_M$ and let E_x denote the set of all edges incident on x in $G(V')$. By α-PTI we get:

$$W(E_x) \geq \alpha W(M). \tag{9}$$

Let's again sum up the previous Inequality 8 over all the edges E_i, $1 \leq i \leq \lfloor p/2 \rfloor$. On the left-hand side, each edge in $G(V')$ occurrs twice, except the edges in M (that do not occur at all) and the edges in E_x that occur once, each. Thus, $2[W(V') - W(M)] - W(E_x) \geq \alpha(p-2)W(M)$. Now, applying the inequality 9, we obtain $2[W(V') - W(M)] \geq \alpha(p-1)W(M)$. If we now substitute $W(V') = W'(V')p(p-1)/2$ and $W(M) = W'(M)(p-1)/2$ and divide both sides by $(p-1)/2$ we will quickly obtain that $W'(V') \geq (\alpha/2)W'(M)[p-1+(2/\alpha)]/p$ that is equivalent to $W'(V') > (\alpha/2)W'(M)$ (for $0 < \alpha < 2$). This completes the Case 2 and the whole proof of the Lemma. *(Quod erat demonstrandum)*$\diamondsuit$

The following proof of Theorem 1 is an extension of the one proposed in [13] (and later presented in [12]) by properly introducing the parameter α.

Proof. (of Theorem 1 from Section 5.1) Let P^* and P be the set of nodes in an optimal solution and that in the solution returned by the HRT algorithm for instance I, respectively. By definition, $OPT(I) = W'(P^*)$ and $HRT(I) = W'(P)$. Let M^* and M denote a maximum-weight $\lfloor p/2 \rfloor$-matching in P^* and in P, respectively. By Lemma 3, we get:

$$OPT(I) \leq W'(M^*). \tag{10}$$

In addition, from Lemma 4 we get:

$$HRT(I) > (\alpha/2)W'(M). \tag{11}$$

Now, because the algorithm HRT finds a maximum-weight $\lfloor p/2 \rfloor$-matching in G, we get $W'(M) \geq W'(M^*)$. This, together with the inequalities 10 and 11 implies that $HRT(I) > (\alpha/2)W'(M) \geq (\alpha/2)W'(M^*) \geq OPT(I)/\frac{2}{\alpha}$ that completes the proof of the theorem. *(Quod erat demonstrandum)*$\diamondsuit$

8.2 Proof of Theorem 2 from Section 5.3

The proof constitutes an extension of the one presented in [15] by properly introducing the parameter α.

Let define $\lambda = \frac{2}{\alpha}$, where α is the value of the parameter in α-PTI satisfied by the distance function d. Let P denote the set-valued variable used in GMM presented in Figure 2. By induction on size of P we will show the following condition:

$$f_{MIN}(P) \geq OPT(I)/\lambda \tag{12}$$

holds after each addition to P. After the last addition to P, $GMM(I) = f_{MIN}(P)$ holds, that would imply the theorem.

Initially, the condition holds due to adding two vertices joined by the heaviest edge in V to P. Let's make the inductive assumption that the condition holds after k additions to P, for some $1 \leq k < p-1$ (notice that after k additions P contains $k+1$ elements). It will be proven that the condition also holds after the $(k+1)$-th addition to P.

Let $P^* = \{v_1^*, \ldots, v_p^*\}$ denote an optimal solution to I and let $l^* = OPT(I)$.

Observation 1: $d(v_i^*, v_j^*) \geq l^*$ for any $i \neq j$, due to the minimality of l^*.

Let $P_k = \{x_1, x_2, \ldots, x_{k+1}\}$ denote the set P after k additions for $1 \leq k < p-1$. Because GMM adds at least one more node to P the following holds:

Observation 2: For $1 \leq k < p-1$, $|P_k| = k+1 < p$.

Let, for every $v_i^* \in P^*$, define $S_i^* = \{u \in V \,|\, d(v_i^*, u) < l^*/\lambda\}$. Notice that for any $1 \leq i \leq p$, $S_i^* \neq \emptyset$, since $v_i^* \in S_i^*$ due to the discernibility property of the distance function d (see Section 4).

Furthermore, for any $1 \leq i < j \leq p$, S_i^* and S_j^* are disjoint. To prove this, assume the opposite, i.e. that $S_i^* \cap S_j^* \neq \emptyset$ for some $i \neq j$. Let $u \in S_i^* \cap S_j^*$. Thus $d(v_i^*, u) < l^*/\lambda$ and $d(v_j^*, u) < l^*/\lambda$. This, together with α-PTI implies $\alpha d(v_i^*, v_j^*) \leq d(v_i^*, u) + d(v_j^*, u) < 2l^*/\lambda = \alpha l^*$ that is equivalent to $d(v_i^*, v_j^*) < l^*$. But this would contradict the Observation 1.

Thus $\mathcal{S} = \{S_i^*\}_{1 \leq i \leq p}$ constitutes a family of p non-empty and pair-wise disjoint sets. Thus, since for $k < p-1$, P_k has strictly less than p elements and there are p nonempty sets in family $\mathcal{S}$ there must be at least one index r, $1 \leq r \leq p$ such that $S_r^* \cap P_k = \emptyset$.

Due to the definition of S_r^*, for each $u \in P_k$, $d(v_r^*, u) \geq l^*/\lambda$. Due to the fact that v_r^* is available for selection in the $k+1$-th step of GMM and GMM will select a vertex $v \in V \setminus P_k$ that maximises $min_{v' \in P_k} d(v, v')$ among all the vertices in $V \setminus P_k$ it is implied that the condition 12 still holds after the $(k+1)$-th addition to P (having made the inductive assumption). *(Quod erat demonstrandum)*$\diamondsuit$