Cycle-based formulations in Distance Geometry

Leo Liberti, Gabriele Iommazzo, Carlile Lavor, Nelson Maculan

OJMO

Open Journal of Mathematical Optimization

Leo Liberti, Gabriele Iommazzo, Carlile Lavor & Nelson Maculan

Cycle-based formulations in Distance Geometry

Volume 4 (2023), article no. 1 (16 pages)

https://doi.org/10.5802/ojmo.18

Article submitted on March 16, 2021, revised on August 22, 2022,
accepted on November 9, 2022.

Open Journal of Mathematical Optimization is a member of the
Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/
e-ISSN: 2777-5860

Cycle-based formulations in Distance Geometry

Leo Liberti
LIX CNRS Ecole Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau, France
liberti@lix.polytechnique.fr

Gabriele Iommazzo
Zuse Institute Berlin, Berlin, 14195, Germany
iommazzo@zib.de

Carlile Lavor
IMECC, University of Campinas, Brazil
clavor@ime.unicamp.br

Nelson Maculan
COPPE, Federal University of Rio de Janeiro (UFRJ), Brazil
maculan@cos.ufrj.br

Abstract

The distance geometry problem asks to find a realization of a given simple edge-weighted graph in a Euclidean space of given dimension K, where the edges are realized as straight segments of lengths equal (or as close as possible) to the edge weights. The problem is often modelled as a mathematical programming formulation involving decision variables that determine the position of the vertices in the given Euclidean space. Solution algorithms are generally constructed using local or global nonlinear optimization techniques. We present a new modelling technique for this problem where, instead of deciding vertex positions, the formulations decide the length of the segments representing the edges in each cycle in the graph, projected in every dimension. We propose an exact formulation and a relaxation based on a Eulerian cycle. We then compare computational results from protein conformation instances obtained with stochastic global optimization techniques on the new cycle-based formulation and on the existing edge-based formulation. While edge-based formulations take less time to reach termination, cycle-based formulations are generally better on solution quality measures.

Digital Object Identifier 10.5802/ojmo.18

2020 Mathematics Subject Classification 90C26, 51K05.

Keywords Mathematical Programming, cycle basis, protein conformation.

Acknowledgments While the seminal idea for considering DGPs over cycles dates from Saxe's **NP**-hardness proof [50], the "cycle formulation" concept occurred to us as one of the authors (LL) attended a talk by Matteo Gallet given at the Erwin Schrödinger Institute (ESI), Vienna, during the Geometric Rigidity workshop 2018. LL has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska–Curie grant agreement n. 764759 "MINOA", and from the ANR PRCI project "MultiBioStruct". Most of the work on this paper was carried out while GI was a Ph.D. student at LIX, École Polytechnique. CL is grateful to the Brazilian research agencies FAPESP and CNPq for support. NM is grateful to the Brazilian research agencies COPPETEC Foundation and CNPq for support.

1 Introduction

We consider the fundamental problem in Distance Geometry (DG):

Distance Geometry Problem (DGP). Given a positive integer K and a simple undirected graph $G = (V, E)$ with an edge weight function $d : E \to \mathbb{R}_{\geq 0}$, establish whether there exists a *realization* $x : V \to \mathbb{R}^K$ of the vertices such that Eq. (1) below is satisfied:

$$\forall \{i, j\} \in E \qquad \|x_i - x_j\| = d_{ij}, \tag{1}$$

where $x_i \in \mathbb{R}^K$ for each $i \in V$ and d_{ij} is the weight on edge $\{i, j\} \in E$.

Although the DGP is given above in the canonical decision form, we consider the corresponding search problem, where one has to actually find the realization x. The DGP is also known as the *graph realization problem* in geometric rigidity [6, 17, 28]. It belongs to a more general class of metric completion and embedding problems [7, 23, 51].

In its most general form, the DGP might be parametrized over any norm [11]. In practice, the ℓ_2 norm is the most usual choice [39], and will also be employed in this paper. The DGP with the ℓ_2 norm is sometimes called the Euclidean DGP (EDGP). For the EDGP, Eq. (1) is often reformulated to:

$$\forall \{i,j\} \in E \qquad \|x_i - x_j\|_2^2 = d_{ij}^2, \tag{2}$$

which is a system of quadratic polynomial equations with no linear terms [35, §2.4].

The EDGP is motivated by many scientific and technological applications. The clock synchronization problem, for example, aims at establishing the absolute time of a set of clocks when only the time difference between subsets of clocks can be exchanged [53]. The sensor network localization problem aims at finding the positions of moving wireless sensor on a 2D manifold given an estimation of some of the pairwise Euclidean distances [2, 15, 17]. The Molecular DGP (MDGP) aims at finding the positions of atoms in a protein, given some of the pairwise Euclidean distances [35, 39]. The position of autonomous underwater vehicles cannot be determined via GPS (since the GPS signal does not reach under water), but must rely on distances estimated using sonars: a DGP can then be solved in order to localize the fleet [3]. Applications of the DGP to data science are described in [33]; see [32] for an application to natural language processing. In general, the DGP is an inverse problem that occurs every time one can measure some of the pairwise distances in a set of entities, and needs to establish their position.

The DGP is weakly **NP**-hard even when restricted to simple cycle graphs (by reduction from Partition) and strongly **NP**-hard even when restricted to integer edge weights in $\{1,2\}$ in general graphs (by reduction from 3sat) [50]. It is in **NP** if $K = 1$ but not known to be in **NP** if $K > 1$ for general graphs [4], which is an interesting open question [36].

There are many approaches to solving the DGP. Generally speaking, application-specific solution algorithms exploit some of the graph structure, whenever it is induced by the application. For example, a condition often asked when reconstructing the positions of sensor networks is that the realization should be unique (as one would not know how to choose between multiple realizations), a condition called *global rigidity* [10]. This condition can, at least generically, be ensured by a specific graph rigidity structure of the unweighted input graph, as shown in [20]. For protein structures, on the other hand, which are found in nature in several isomers, one is sometimes interested in finding all (incongruent) realizations of the given protein graph [30, 37, 47]. Since such graphs are rigid, one can devise an algorithm (called Branch-and-Prune) that, following a given vertex order, branches on reflections of the position of the next vertex, which is computed using trilateration [35]. It is also possible that DGP problems arise in their full generality, i.e. independently of any further knowledge on their structure or properties: for such cases, one can resort to Mathematical Programming (MP) formulations and corresponding solvers [12, 14, 40].

The MP formulation that is most often used reformulates Eq. (2) to the minimization of the sum of squared error terms:

$$\min_x \sum_{\{i,j\} \in E} (\|x_i - x_j\|_2^2 - d_{ij}^2)^2. \tag{3}$$

This formulation describes an unconstrained polynomial minimization problem. The polynomial in question has degree 4, is always nonnegative, and generally nonconvex and multimodal. The decision variables are represented by a $n \times K$ rectangular matrix x such that x_{ik} is the k-th component of the vector x_i, which gives the position in $\mathbb{R}^K$ of vertex $i \in V$. Each solution $x^* \in \mathbb{R}^{nK}$ having global minimum value equal to zero is a realization of the given graph. Solutions with small objective function value represent approximate solutions. Because of the nonconvexity of the formulation and the hardness of the problem, Eq. (3) is not usually solved to guaranteed ε-optimality (e.g. using a spatial Branch-and-Bound approach [5]); rather, heuristic approaches, such as MultiStart (MS) [29], Variable Neighbourhood Search (VNS) [38], or relaxation-based heuristics [14, 42] may be used.

As far as we know, all existing MP formulations for the EDGP are edge-based, such as the one in Eq. (3). In this paper we discuss a new MP formulation for the EDGP based on the incidence of cycles and edges instead, a relaxation based on Eulerian cycles, and a computational comparison with Eq. (3).

Although this paper is not about graph theory, a fair amount of graph theoretical content is needed to prove the main reformulation result. The results that follow are presented in a way that does not require much background in graph theory. The paper is self-contained in this respect.

2 Some existing MP formulations

In this short section we give a minimal list of typical variants of Eq. (3) in order to motivate the claim that the cycle-based formulation of the DGP discussed in this paper is new. Of course, only a complete enumeration of DGP formulations in the literature could substantiate this claim. But even this short list shows that the typical modelling approach for the DGP is direct: namely, decision variables encode the realization of each vertex as a vector in $\mathbb{R}^K$. Many more formulations of the DGP and its variants, all corresponding to this criterion, are given in [12, 29, 40].

The closest variant of Eq. (3) simply adds a constraint ensuring that the centroid of all of the points in the realization is at the origin (see Lemma 16 below). This removes the degrees of freedom given by translations:

$$\left.\begin{array}{rl} \min_x & \sum_{\{i,j\}\in E}(\|x_i - x_j\|_2^2 - d_{ij}^2)^2 \\ \forall\, k \leq K & \sum_{i\in V} x_{ik} = 0. \end{array}\right\} \tag{4}$$

This formulation describes a linearly constrained polynomial minimization problem. Like Eq. (3), the polynomial in Eq. (4) has degree 4, is always nonnegative, and is generally nonconvex and multimodal.

Another small variant of Eq. (4) is achieved by adding range bounds to the realization variables x; generally valid (but slack) bound values can be set to $\pm\frac{1}{2}\sum_{\{i,j\}\in E} d_{ij}$. This corresponds to the worst case of a single path being arranged in a straight line with unknown orientation.

Another possible formulation, derived again from Eq. (3), is obtained by replacing the squared error with absolute value errors (whose positive and negative parts are encoded by s^+, s^-). This yields the following formulation:

$$\left.\begin{array}{rl} \min_{s,x} & \sum_{\{i,j\}\in E}(s_{ij}^+ + s_{ij}^-) \\ \forall\, \{i,j\} \in E & \|x_i - x_j\|_2^2 = d_{ij}^2 + s_{ij}^+ - s_{ij}^- \\ \forall\, \{i,j\} \in E & s_{ij}^+, s_{ij}^- \geq 0. \end{array}\right\} \tag{5}$$

Note that, again, each solution s^*, x^* with zero optimal objective value makes x^* an encoding of a realization of the given graph. Thus, global optima are preserved by this reformulation, while local optima may differ.

Yet another reformulation derived from replacing squared errors with absolute values consists in observing that the "plus" and "minus" parts of each absolute value term correspond to a convex and concave function. This yields a formulation called *push-and-pull*, since the objective pulls adjacent vertices apart, while the constraint push them back together:

$$\left.\begin{array}{rlcl} \max_x & \sum_{\{i,j\}\in E} \|x_i - x_j\|_2^2 & & \\ \forall\, \{i,j\} \in E & \|x_i - x_j\|_2^2 & \leq & d_{ij}^2. \end{array}\right\} \tag{6}$$

Eq. (6) is a Quadratically Constrained Quadratic Program with concave objective and convex constraints. It was used within a Multiplicative Weights Update algorithm for the DGP in [12], as well as a basis for Semidefinite Programming and Diagonally Dominant Programming relaxations [14, 42]. It can be shown that all constraints are active at global optima, which therefore correspond to realizations of the given graph [46].

3 A new formulation based on cycles

In this section we propose a new formulation for the EDGP, based on the fact that the quantities $x_{ik} - x_{jk}$ sum up to zero over all edges of any cycle in the given graph for each dimensional index $k \leq K$. This idea was used in [50] for proving weak **NP**-hardness of the DGP on cycle graphs. For a subgraph H of a graph $G = (V, E)$, we use $V(H)$ and $E(H)$ to denote vertex and edge set of H explicitly; given a set F of edges we use $V(F)$ to denote the set of incident vertices. Let $m = |E|$ and $n = |V|$. For a mapping $x : V \to \mathbb{R}^K$ we denote by $x[U]$ the restriction of x to a subset $U \subseteq V$. Furthermore, we let a *closed trail* be a sequence of vertices and of the edges joining them, which begins and ends at the same vertex, and is such that no edge is repeated.

▶ **Lemma 1.** *Given an integer $K > 0$, a simple undirected weighted graph $G = (V, E, d)$ and a mapping $x : V \to \mathbb{R}^K$, then for each cycle C in G, each orientation of the edges in C given by a closed trail $W(C)$ in the cycle, and each $k \leq K$ we have:*

$$\sum_{(i,j)\in W(C)} (x_{ik} - x_{jk}) = 0. \tag{7}$$

Proof. We renumber the vertices in $V(C)$ to $1, 2, \ldots, \gamma = |V(C)|$ following the walk order in $W(C)$. Then Eq. (7) can be explicitly written as:

$$(x_{1k} - x_{2k}) + (x_{2k} - x_{3k}) + \cdots + (x_{\gamma k} - x_{1k}) = x_{1k} - (x_{2k} - x_{2k}) - \cdots - (x_{\gamma k} - x_{\gamma k}) - x_{1k} = 0,$$

as claimed. ◀

We introduce new decision variables y_{ijk} replacing the terms $x_{ik} - x_{jk}$ for each $\{i, j\} \in E$ and $k \leq K$. Eq. (2) then becomes:

$$\forall\, \{i, j\} \in E \qquad \sum_{k\leq K} y_{ijk}^2 = d_{ij}^2. \tag{8}$$

We note that, with a slight abuse of notation, we index the sum in Eq. (8) with the shorthand $k \leq K$ instead of $k \in \{1, 2, \ldots, K\}$. We will keep this notation throughout the paper, for ease of reading. Moreover, we remark that for the DGP with other norms this constraint changes. For the ℓ_1 or ℓ_∞ norms, for example, we would have:

$$\forall\, \{i, j\} \in E \quad \sum_{k\leq K} |y_{ijk}| = d_{ij} \quad \text{or} \quad \max_{k\leq K} |y_{ijk}| = d_{ij}. \tag{9}$$

Next, we adjoin the constraints on cycles:

$$\forall\, k \leq K, C \subseteq E \quad \left(C \text{ is a cycle} \Rightarrow \sum_{\{i,j\}\in E(C)} y_{ijk} = 0\right). \tag{10}$$

We also note that the feasible value of a y_{ijk} variable is the (oriented) length of the segment representing the edge $\{i, j\}$ projected on the k-th coordinate. We can therefore infer bounds for y as follows:

$$\forall\, k \leq K, \{i, j\} \in E \quad -d_{ij} \leq y_{ijk} \leq d_{ij}. \tag{11}$$

Although Eq. (11) are not necessary to solve the cycle formulation, they may improve performance of spatial Branch-and-Bound (sBB) algorithms [5, 54] and of various "matheuristics" [41] that need explicit bounds on all variables, as well as allow an exact linearization of variable products, should a y variable occur in a product with a binary variable in some DGP variant.

We now give the following definition and state our main result, i.e., that Eq. (8) and (10) are a valid MP formulation for the EDGP.

▶ **Definition 2.** *Given a strictly positive $K \in \mathbb{N}$ and a graph $G = (V, E)$, $Y \triangleq \{y \in \mathbb{R}^{Km} \mid (8) \wedge (10)\}$ is the set of vectors satisfying Eq.* (8) *and* (10)*.*

We emphasize that Y depends on the EDGP instance (K, G).

▶ **Theorem 3.** *The set Y is non-empty if and only if (K, G) is a YES instance of the EDGP.*

The proof argues by recursion on a graph decomposition of G that a certain linear system related to the cycles of G (see Eq. (12) below) has a solution in the x variables if and only if the given EDGP instance is YES, as certified by the y variables[1].

We shall construct our proof by steps. The first step defines a graph decomposition based on the removal of a single vertex. Given a graph $G = (V, E)$ and a subset $U \subset V$, the subgraph $G[U]$ *induced* by U is the graph $(U, \{\{u, v\} \in E \mid u, v \in U\})$. With a slight abuse of notation we denote the vertices of a graph G' by $V(G')$ and

[1] This is not the only way to construct x from y: three colleagues, in three separate occasions, have suggested that path lengths (as measured by sums of y variables) can yield valid values for the x variables in each dimension: then, the cycle condition would prove consistency of x and y. This is easy enough to explain informally. When we set about formalizing this suggestion, so that it would be clear in all its parts, we realized that the proof would likely be as long as the one we present here.

its edges by $E(G')$. We let $\gamma(G)$ be the number of connected components of G. A vertex v of G with the property that $\gamma(G[V \smallsetminus \{v\}]) > \gamma(G)$ is called a *cut vertex*. A graph G is *biconnected* if, for any pair u, v of distinct vertices of G, there is a simple cycle in G incident to u and v. It is not hard to show that biconnectedness is equivalent to connectedness and the absence of cut vertices. To see this, we first introduce the concept of "1-decomposition", then prove some statements related to it.

▶ **Definition 4.** *A* 1-decomposition *of a graph $G = (V, E)$ is a set of subgraphs $G_1, \ldots, G_r$ (where $r \in \mathbb{N}$ with $r \geq 1$) of G such that:*
(a) *G_i is either biconnected or a tree for all $i \leq r$;*
(b) *$\bigcup_{i \leq r} E(G_i) = E$;*
(c) *for any $i < j \leq r$ the intersection $V(G_i) \cap V(G_j)$ is either empty or it consists of a single cut vertex of G.*
A 1-decomposition of G is nontrivial *if $r > 1$. A graph G is 1-decomposable if it has a nontrivial 1-decomposition.*

The 1-decomposition bears some relationship to the block-cutpoint tree defined by Harary in [22, p. 36]. However, subgraphs in the 1-decomposition may also be trees, which cannot appear in Harary's construction, since every vertex of a tree is a cutpoint by definition. Trees are important because they are easy to realize in $\mathbb{R}^K$. Their realizations can then be pasted to the realizations of the other subgraphs by rotations and translations, a fact that is used in the proof of the main theorem. The same would not follow if we were to use Harary's block-cutpoint trees, since they contract blocks to a single vertex. We do, however, invoke [22, Thm. 3.1] to state that a connected graph $G = (V, E)$ is 1-decomposable if and only if it has a cut vertex.

▶ **Lemma 5.** *Let G be 1-decomposable, with decomposition $\mathcal{G} = \{G_1, \ldots, G_r\}$, and C be a cycle in G. Then there is an index $i \leq r$ s.t. C is a subgraph of G_i.*

Proof. Suppose, to aim at a contradiction, that there are two distinct subgraphs G_i, G_j in $\mathcal{G}$ both incident to the edges of C. Then there is a nontrivial path p in C, with at least two edges, joining a vertex u in G_i to a vertex v in G_j. Therefore, by [22, Thm. 3.1], there must be a cut vertex of G on p, which implies that there is a cut vertex in C, which is impossible, since cycles are biconnected. ◀

We note that no biconnected graph G is 1-decomposable. On the other hand, a tree with n vertices can always be 1-decomposed into n subgraphs.

▶ **Proposition 6.** *Any connected component $G = (V, E)$ of a simple graph has a (possibly trivial) 1-decomposition consisting of biconnected subgraphs and tree subgraphs.*

Proof. We prove this result by induction on the number β of biconnected subgraphs in a 1-decomposition $\mathcal{C} = \{G_1, \ldots, G_r\}$ of G for some $r \in \mathbb{N}$. We first deal with the base case, where $\beta = 0$. We claim that G must be a tree: supposing G has a cycle G', as well as biconnectedness of cycles and part (c) of Definition 4, G' must be one of the $G_1, \ldots, G_r$. But then $\beta \geq 1$ against the assumption. Therefore, the trivial 1-decomposition $\mathcal{C} = \{G\}$ is a valid 1-decomposition of G. We now tackle the induction step. Consider the largest biconnected subgraph B of G: then $\widetilde{G} = G[V \smallsetminus V(B)]$ has one fewer biconnected components than G, so, by induction, $\widetilde{G}$ has a 1-decomposition $\mathcal{D}' = \{G'_1, \ldots, G'_{t-1}\}$ for some $t \in \mathbb{N}$ with $t > 1$. We prove that $\mathcal{D} = \mathcal{D}' \cup \{B\}$ is a valid 1-decomposition of G. Condition (a) is verified since $\mathcal{D}'$ is a valid 1-decomposition by induction, and B is biconnected; condition (b) is verified since the union of the graph in $\mathcal{D}$ is $\mathcal{G}$ by construction; for condition (c), suppose there is $i < t$ s.t. $|V(G_i) \cap V(B)| \geq 2$: this means there are two distinct vertices u, v in both $V(G_i)$ and $V(B)$. Since G_i is connected, there must be a path p from u to v in G_i, hence $G[B \cup V(p)]$ is a biconnected graph larger than B. But B was assumed to be largest, so this is not possible, and (c) holds, which concludes the proof. ◀

The second step proves the easier ($\Leftarrow$) direction of Theorem 3.

▶ **Proposition 7.** *For any YES instance (K, G) of the EDGP there is a vector $y^* \in Y$.*

Proof. Assume that (K, G) is a YES instance of the EDGP. Then G has a realization $x^* \in \mathbb{R}^{nK}$ in $\mathbb{R}^K$. We define $y^*_{ijk} = x^*_{ik} - x^*_{jk}$ for all $\{i, j\} \in E$ and $k \leq K$. Since x^* is a realization of G, by definition it satisfies Eq. (2), and, by substitution, Eq. (8). Moreover, any realization of G satisfies Eq. (7) over each cycle by Lemma 1. Hence, by replacement, it also satisfies Eq. (10). ◀

In the third step, we lay the groundwork towards the more difficult ($\Rightarrow$) direction of Theorem 3. We proceed by contradiction: we assume that (K, G) is a NO instance of the EDGP, and suppose that the set Y for this

instance is non-empty. For every $y \in Y$ we consider the K linear systems

$$\forall \{i,j\} \in E \quad x_{ik} - x_{jk} = y_{ijk}, \tag{12}$$

for each $k \leq K$, each with n variables and m equations. We square both sides then sum over $k \leq K$ to obtain

$$\forall \{i,j\} \in E \quad \sum_{k \leq K} (x_{ik} - x_{jk})^2 = \sum_{k \leq K} y_{ijk}^2. \tag{13}$$

By Eq. (8) we have

$$\sum_{k \leq K} y_{ijk}^2 = d_{ij}^2, \tag{14}$$

whence follows Eq. (2), contradicting the assumption that the EDGP is NO. So we only need to show that there is a solution x^* to Eq. (12) for any given $y \in Y$. To this effect, we shall exploit the 1-decomposition of G into biconnected graphs and trees derived in Proposition 6. First, though, we have to show that Eq. (12) has a solution if $Y \neq \varnothing$ in the "base cases" of the 1-decomposition, namely trees and biconnected graphs.

The following result essentially proves that the constraint matrix of Eq. (12) has full rank, which is an easy consequence of graphic matroid theory. We prove the result by elementary means for self-containment.

▶ **Lemma 8.** *Let $G = (V, E)$ be a tree, and $Y \neq \varnothing$. Then Eq. (12) has a solution for every $k \leq K$.*

Proof. Let M be the coefficient matrix of the system of equations (12), for a given $k \leq K$; and let y^k be the vector $(y_{uvk} \mid \{u,v\} \in E)$. We note that, since M is the (transposed) incidence matrix of G, only the right-hand side of the system changes for each k. We aim at proving that M and (M, y^k) have the same rank, and that this rank is full. We proceed by induction on the size $|E|$ of the tree. The base case, where $|E| = 1$ and G consists of a single edge $\{u,v\}$, yields $M = (1, -1)$ with rank 1 for each $k \leq K$. By inspection, (M, y_{uvk}) also has rank 1 for any y_{uvk}. Consider a tree G' with one fewer edge (say, $\{u,v\}$) than G, such that $V \smallsetminus V(G') = \{v\}$. Let the corresponding system Eq. (12) $\widetilde{M}x = \widetilde{y}$ satisfy $\mathsf{rank}(\widetilde{M}) = \mathsf{rank}(\widetilde{M}, \widetilde{y}^k)$, for all $k \leq K$. Then the shape of M is:

$$M = \begin{pmatrix} \widetilde{M} & 0 \\ e_u & -1 \end{pmatrix},$$

where $e_u = (0, \ldots, 0, 1_u, 0, \ldots, 0)$. This shows that $\mathsf{rank}(M) = \mathsf{rank}(\widetilde{M}) + 1$, that this rank is full, and hence also that $\mathsf{rank}(M) = \mathsf{rank}((M, y^k))$. ◀

▶ **Lemma 9.** *Let $G = (V, E)$ be biconnected, and $Y \neq \varnothing$. Then Eq. (12) has a solution for every $k \leq K$.*

Proof. We proceed by induction on the simple cycles of G. For the base case, we consider G to be a graph consisting of a single cycle, with corresponding $y \in Y$. Since G is a cycle, it has the same number of vertices and edges, say q. This implies that, for any fixed $k \leq K$, Eq. (12) is a linear system $Mx = y^k$ (where $y^k = (y_{uvk} \mid \{u,v\} \in E)$) with a $q \times q$ coefficient matrix:

$$M = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & \ddots & & \\ & & & \ddots & -1 & \\ -1 & & & & 1 & \end{pmatrix}. \tag{15}$$

We remark that M is the incidence matrix of G as in the Proof of Lemma 8. By Eq. (7) and by inspection of Eq. (15) it is clear that $\mathsf{rank}(M) = q - 1$: then Eq. (10) ensures that $\mathsf{rank}((M, y^k)) = \mathsf{rank}(M)$, and therefore that Eq. (12) has a solution.

We now tackle the induction step. The incidence vectors in E of the cycles of any graph are a vector space of dimension $m - n + 1$ over the finite field $\mathbb{F}_2 = \{0, 1\}$ [52]. We consider a fundamental cycle basis $\mathcal{B}$ of G (see Section 4). We assume that (a) G' is a union of fundamental cycles in $\mathcal{B}' \subsetneq \mathcal{B}$, for which Eq. (12) has a solution x' by the induction hypothesis, and (b) that C is another fundamental cycle in $\mathcal{B} \smallsetminus \mathcal{B}'$, with a solution x^C of Eq. (12) that exists by the base case. We aim at proving that Eq. (12) has a solution for $G' \cup C$. Since G is

biconnected, the induction can proceed by ear decomposition [44], which means that G' is also biconnected, and that C is such that $E(G') \cap E(C) = F$ is a non-empty path in G'.

By Eq. (10) applied to C, we have

$$\forall\, k \leq K \quad \sum_{\{i,j\} \in C} y_{ijk} = 0. \tag{16}$$

Since x' satisfies Eq. (12) by the induction hypothesis,

$$\forall\, k \leq K, \{i,j\} \in F \quad x'_{ik} - x'_{jk} = y_{ijk}. \tag{17}$$

We replace Eq. (17) in Eq. (16), obtaining

$$\forall\, k \leq K \quad \sum_{\{i,j\} \in F} (x'_{ik} - x'_{jk}) = - \sum_{\{i,j\} \in E(C) \smallsetminus F} y_{ijk}. \tag{18}$$

Moreover, x^C also satisfies Eq. (12) over C, hence we can replace the right hand side of Eq. (18) with the corresponding terms in $x^C_{ik} - x^C_{jk}$ to get:

$$\forall\, k \leq K \quad \sum_{\{i,j\} \in F} (x'_{ik} - x'_{jk}) + \sum_{\{i,j\} \in E(C) \smallsetminus F} (x^C_{ik} - x^C_{jk}) = 0. \tag{19}$$

We now fix x', and aim at modifying x^C so that: (a) x^C matches x' on $V(F)$, (b) the modified x^C is still a solution of Eq. (12) on C. We set x^C_{ik} to x'_{ik} for each $i \in V(F)$, and consider the resulting linear system Eq. (12) given by M, as in Eq. (15), for each $k \leq K$, where we assume without loss of generality that $V(F) = \{1, \ldots, r\}$ and $V(C) = \{r+1, \ldots, s\}$:

$$\left.\begin{array}{rrrrrrrrrrrrl}
x'_{1k} & - x'_{2k} & & & & & & & & = & y_{12k} & (1) \\
 & x'_{2k} & - x'_{3k} & & & & & & & = & y_{23k} & (2) \\
 & \ddots & \ddots & & & & & & & \vdots & \vdots & \vdots \\
 & & x'_{rk} & - x^C_{r+1,k} & & & & & & = & y_{r,r+1,k} & (r) \\
 & & & x^C_{r+1,k} & - x^C_{r+2,k} & & & & & = & y_{r+1,r+2,k} & (r{+}1) \\
 & & & & \ddots & \ddots & & & & \vdots & \vdots & \vdots \\
 & & & & & x^C_{s-1,k} & - x^C_{sk} & & & = & y_{s-1,s,k} & (s{-}1) \\
- x'_{1k} & & & & & & & & x^C_{sk} & = & y_{1sk}. & (s)
\end{array}\right\} \tag{20}$$

The equations from (1) to $(r-1)$ in Eq. (20) are satisfied by the induction hypothesis since they only depend on x', so we can remove them from the system and assume x' to be constant. We are left with:

$$\left.\begin{array}{rrrrrl}
- x^C_{r+1,k} & & & = & y_{r,r+1,k} - x'_{rk} & (r) \\
x^C_{r+1,k} & - x^C_{r+2,k} & & = & y_{r+1,r+2,k} & (r{+}1) \\
\ddots & \ddots & & \vdots & \vdots & \vdots \\
 & x^C_{s-1,k} & - x^C_{sk} & = & y_{s-1,s,k} & (s{-}1) \\
 & & x^C_{sk} & = & y_{1sk} + x'_{1k}. & (s)
\end{array}\right\} \tag{21}$$

Summing up the left hand sides of Eq. (21), we obtain:

$$\begin{aligned}
&- x^C_{r+1,k} + (x^C_{r+1,k} - x^C_{r+2,k}) + \cdots + (x^C_{s-1,k} - x^C_{sk}) + x^C_{sk} \\
&= (-x^C_{r+1,k} + x^C_{r+1,k}) + \cdots + (-x^C_{sk} + x^C_{sk}) = 0
\end{aligned}$$

for all $k \leq K$, so the $(s-r+1) \times (s-r+1)$ matrix $\overline{M}$ of the k-th linear system Eq. (21) has rank $\leq s-r$. On the other hand, eliminating the first or last row makes it clear by inspection that the rest of the rows are linearly independent; therefore the rank of $\overline{M}$ is exactly $s-r$. Summing up the components of the right hand side vector $\overline{y}^k$ of Eq. (21), we obtain:

$$\begin{aligned}
\chi &= -x'_{rk} + y_{r,r+1,k} + y_{r+1,r+2,k} + \cdots + y_{s-1,s,k} + y_{1sk} + x'_{1k} \\
&= (x'_{1k} - x'_{rk}) + \sum_{\{i,j\} \in E(C) \smallsetminus F} y_{ijk}.
\end{aligned}$$

We remark that

$$\begin{aligned} x'_{1k} - x'_{rk} &= (x'_{1k} - x'_{2k}) + (x'_{2k} - x'_{3k}) + \cdots + (x'_{r-1,k} + x'_{rk}) \\ &= \sum_{\{i,j\}\in F} (x'_{ik} - x'_{jk}) = \sum_{\{i,j\}\in F} y_{ijk} \end{aligned}$$

since x' satisfies Eq. (12) by the induction hypothesis. Therefore

$$\chi = \sum_{\{i,j\}\in F} y_{ijk} + \sum_{\{i,j\}\in E(C)\smallsetminus F} y_{ijk} = \sum_{\{i,j\}\in E(C)} y_{ijk},$$

whence $\chi = 0$ by Eq. (16). This implies that $\mathsf{rank}((\overline{M}, \bar{y}^k)) = \mathsf{rank}(\overline{M}) = s - r$. Therefore, Eq. (21) has a solution, which yields the modified x^C with properties (a) and (b) given above. This concludes the induction step and the proof. ◀

We can finally give the proof of Theorem 3.

Proof of Theorem 3. The ($\Leftarrow$) part follows by Proposition 7. For the ($\Rightarrow$) part, we exploit a 1-decomposition of G into trees and biconnected subgraphs, derive solutions to Eq. (12) for each subgraph, and show that the solutions can be easily combined to yield a solution to Eq. (12) for the whole graph G.

We assume without loss of generality that G is connected (otherwise each connected component can be treated separately), and consider a 1-decomposition $\mathcal{D} = \{G_1, \ldots, G_r\}$ of G. By Lemmata 8 and 9, there exist solutions $x^1, \ldots, x^r$ to Eq. (12) applied to $G_1, \ldots, G_r$ respectively. Consider the graph

$$\mathcal{D} = (\mathcal{D},\ \{\{i,j\} \mid 1 \leq i \neq j \leq r \wedge |V(G_i) \cap V(G_j)| = 1\}).$$

By Lemma 5, $\mathcal{D}$ is a tree: otherwise, a cycle in $\mathcal{D}$ would be a contraction of a cycle in G not included in a single G_i, against Lemma 5. This allows us to reorder $\mathcal{D}$ so that, for each $j > 1$, there is a unique $i < j$ such that $\{i,j\} \in E(\mathcal{D})$.

We remark that, for each $i \leq r$, x^i is a realization of G_i in $\mathbb{R}^K$ by Eq. (12)–(14). More precisely, x^i is a $|V(G_i)| \times K$ matrix $x^i = (x^i_{\ell k})$ so that $x^i_\ell = (x^i_{\ell 1}, \ldots, x^i_{\ell K})$ is the position of vertex $\ell \in V(G_i)$ in $\mathbb{R}^K$. Note that the realizations $x^1, \ldots, x^r$ can be modified by translations without changing the values of y (by inspection of Eq. (12)).

We now construct a solution $\bar{x}$ of Eq. (12) for G by induction on $\mathcal{D}$ ordered as described above. For the base case $i = 1$, we fix x^1 in any way (e.g. by taking the centroid of the rows of x^1 to be the origin), and initialize the first $|V(G_1)|$ rows of $\bar{x}$ with those of x^1. For any $i > 1$, we identify the unique predecessor j of i in the order on $\mathcal{D}$. The induction hypothesis ensures the existence of a solution $\bar{x}$ of the union of $G_1, \ldots, G_j$. Consider the cut vertex v in $V(G_j) \cap V(G_i)$ guaranteed by definition of the order on $\mathcal{D}$, and let $\bar{x}_v \in \mathbb{R}^K$ be its position. Then the translation $\widetilde{x}^i = x^i - \mathbf{1}(x^i_v - \bar{x}_v)^\top$ yields another valid solution of Eq. (12) applied to G^i by translation invariance, and this solution is such that $\widetilde{x}^i_v = \bar{x}_v$. Therefore, using the rows of $\widetilde{x}^i$, $\bar{x}$ can be extended to a solution of Eq. (12) applied to the union of $G_1, \ldots, G_j$ and G^i, as claimed. ◀

Theorem 3 can also be interpreted as a polynomial reduction of the EDGP to the problem of finding a solution of Eq. (8) and (10).

▶ **Corollary 10.** *Deciding feasibility of Eq.* (8) *and* (10) *is* **NP***-hard.*

Proof. By reduction from EDGP using Theorem 3. ◀

A remarkable consequence of Theorem 3 is that it allows a decomposition of the computation of the realization x into two stages: first, solve Eq. (8)–(10) to find a feasible y^*; then solve

$$\forall\, k \leq K, \{i,j\} \in E \quad x_{ik} - x_{jk} = y^*_{ijk} \tag{22}$$

to find a realization x^*. We note that Eq. (22) is just a restatement of Eq. (12) universally quantified over k.

▶ **Corollary 11.** *Given an EDGP instance (K, G) and a solution $y^* \in Y$, any solution x^* of Eq.* (22) *is a valid realization of the given instance.*

Proof. The feasibility of Eq. (22) with the right hand side replaced by $y^* \in Y$ follows directly from Theorem 3, since if such a y^* exists then the EDGP is feasible. ◀

The first stage is **NP**-hard by Corollary 10, while the second stage is tractable, since solving linear systems can be done in polynomial time.

▶ Remark 12. Note that Eq. (22) has Km equations, but its rank may be lower, since there are only Kn variables: in particular, Eq. (22) may be an overdetermined linear system. The feasibility of this system is guaranteed by Corollary 11; in particular, the steps of the proof of Theorem 3 imply that Eq. (22) loses rank w.r.t. Km according to the incidence of the edges in the cycles of G. In other words, any solution y' to Eq. (10) provides a right hand side to Eq. (22) that makes the system feasible.

The issue with Theorem (3) is that it relies on the exponentially large family of constraints Eq. (10). While this is sometimes addressed by algorithmic techniques such as row generation, we shall see in the following that it suffices to consider a polynomial set of cycles (which, moreover, can be found in polynomial time) in the quantifier of Eq. (10).

4 The cycle vector space and its bases

We recall that incidence vectors of cycles (in a Euclidean space having $|E|$ dimensions) form a vector space over a field $\mathbb{F}$, which means that every cycle can be expressed as a weighted sum of cycles in a basis. In this interpretation, a *cycle* in G is simply a subgraph of G where each vertex has even degree: we denote their set by $\mathcal{C}$. This means that Eq. (10) is actually quantified over a subset of $\mathcal{C}$, namely the simple connected cycles. Every basis has cardinality $m - n + a$, where a is the number of connected components of G. If G is connected, cycle bases have cardinality $m - n + 1$ [52].

Our interest in introducing cycle bases is that we would like to quantify Eq. (10) polynomially rather than exponentially in the size of G. Our goal is to replace "C is any simple connected cycle in $\mathcal{C}$" by "C is a cycle in a cycle basis of G". In order to show that this limited quantification is enough to imply every constraint in Eq. (10), we have to show that, for each simple connected cycle $C \in \mathcal{C}$, the corresponding constraint in Eq. (10) can be obtained as a weighted sum of constraints corresponding to the basis elements.

Another feature of Eq. (10) to keep in mind is that edges are implicitly given a direction: for each cycle, the term for the *undirected* edge $\{i, j\}$ in Eq. (10) is $(x_{ik} - x_{jk})$. Note that while $\{i, j\}$ is exactly the same vertex set as $\{j, i\}$, the corresponding term is either positive or not, depending on the direction (i, j) or (j, i). We deal with this issue by arbitrarily directing the edges in E to obtain a set A of arcs, and considering *directed* cycles in the directed graph $\overline{G} = (V, A)$. In this interpretation, the incidence vector of a directed cycle C of $\overline{G}$ is a vector $c^C \in \mathbb{R}^m$ satisfying [27, §2, p. 201]:

$$\forall\, j \in V(C) \qquad \sum_{(i,j)\in A} c^C_{ij} = \sum_{(j,\ell)\in A} c^C_{j\ell}. \tag{23}$$

A directed circuit D of $\overline{G}$ is obtained by applying the edge directions from $\overline{G}$ to a connected subgraph of G where each vertex has degree exactly 2 (note that a directed circuit need not be strongly connected, although its undirected version is connected). Its incidence vector $c^D \in \{-1, 0, 1\}^m$ is defined as follows:

$$\forall\, (i, j) \in A \qquad c^D_{ij} \triangleq \begin{cases} 1 & \text{if } (i, j) \in A(D) \\ -1 & \text{if } (j, i) \in A(D) \\ 0 & \text{otherwise} \end{cases}$$

where we have used $A(D)$ to mean the arcs in the subgraph D. In other words, whenever we walk over an arc (i, j) in the natural direction $i \to j$ we let the (i, j)-th component of c^D be 1; if we walk over (i, j) in the direction $j \to i$ we assign a -1, and otherwise a zero.

4.1 Constraints over cycle bases

The properties of undirected and directed cycle bases have been investigated in a sequence of papers by many authors, culminating with [27]. We now prove that it suffices to quantify Eq. (10) over a directed cycle basis.

▶ **Proposition 13.** *Let $\mathcal{B}$ be a directed cycle basis of $\overline{G}$ over $\mathbb{Q}$. Then Eq.* (10) *holds if and only if:*

$$\forall\, k \leq K, B \in \mathcal{B} \qquad \sum_{(i,j)\in A(B)} c^B_{ij} y_{ijk} = 0. \tag{24}$$

Proof. Necessity (10) ⇒ (24) follows because Eq. (10) is quantified over all cycles: in particular, it follows for any undirected cycle in any undirected cycle basis. Moreover, the signs of all terms in the sum of Eq. (24) are consistent, by definition, with the arbitrary edge direction chosen for $\overline{G}$.

Next, we claim sufficiency (24) ⇒ (10). Let $C \in \mathcal{C}$ be a simple cycle, and $\overline{C}$ be its directed version with the directions inherited from $\overline{G}$. Since $\mathcal{B}$ is a cycle basis, we know that there is a coefficient vector $(\gamma_B \mid B \in \mathcal{B}) \in \mathbb{R}^{|\mathcal{B}|}$ such that:

$$c^{\overline{C}} = \sum_{B \in \mathcal{B}} \gamma_B c^B. \tag{25}$$

We now consider the expression:

$$\forall\, k \leq K \quad \sum_{B \in \mathcal{B}} \gamma_B \sum_{(i,j) \in A(B)} c^B_{ij} y_{ijk}. \tag{26}$$

On the one hand, by Eq. (25), Eq. (26) is identically equal to $\sum_{(i,j)\in A(\overline{C})} c^{\overline{C}}_{ij} y_{ijk}$ for each $k \leq K$; on the other hand, each inner sum in Eq. (26) is equal to zero by Eq. (24). This implies $\sum_{(i,j)\in A(\overline{C})} c^{\overline{C}}_{ij} y_{ijk} = 0$ for each $k \leq K$. Since C is simple and connected, $\overline{C}$ is a directed circuit. This implies that $c^{\overline{C}} \in \{-1, 0, 1\}$. Now it suffices to replace $-y_{ijk}$ with y_{jik} to obtain

$$\forall\, k \leq K \quad \sum_{\{i,j\} \in E(C)} y_{ijk} = 0,$$

where the edges on C are indexed in such a way as to ensure they appear in order of consecutive adjacency. ◀

Obviously, if $\mathcal{B}$ has minimum (or just small) cardinality, Eq. (24) will be sparsest (or just sparse), which is often a desirable property of linear constraints occurring in MP formulations. Hence we should attempt to find short cycle bases $\mathcal{B}$.

In summary, given a basis $\mathcal{B}$ of the directed cycle space of $\overline{G}$ where c^B is the incidence vector of a cycle $B \in \mathcal{B}$, the following:

$$\left.\begin{array}{rlcl} \min_{s\geq 0, y} & \sum_{\{i,j\}\in E}(s^+_{ij} + s^-_{ij}) & & \\ \forall\,(i,j) \in A(\overline{G}) & \sum_{k\leq K} y^2_{ijk} - d^2_{ij} & = & s^+_{ij} - s^-_{ij} \\ \forall\, k \leq K, B \in \mathcal{B} & \sum_{(i,j)\in A(B)} c^B_{ij} y_{ijk} & = & 0 \end{array}\right\} \tag{27}$$

is a valid formulation for the EDGP. The solution of Eq. (27) yields a feasible vector y^*. As pointed out in Corollary 11, we must then solve Eq. (22) to obtain a realization x^* for G.

4.2 How to find directed cycle bases

We require directed cycle bases over $\mathbb{Q}$. By [27, Thm. 2.4], each undirected cycle basis gives rise to a directed cycle basis (so it suffices to find a cycle basis of G and then direct the cycles using the directions in $\overline{G}$). Horton's algorithm [24] and its variants [19, 43] find a minimum cost cycle basis in polynomial time. The most efficient deterministic variant is $O(m^3 n)$ [43], and the most efficient randomized variant has the complexity of matrix multiplication. Existing approximation algorithms have marginally better complexity.

It is not clear, however, that the provably sparsest constraint system will make the DGP actually easier to solve. We therefore consider a much simpler algorithm: starting from a spanning tree, we pick the $m - n + 1$ circuits that each *chord* (i.e., non-tree) edge defines with the rest of the tree. This algorithm [48] yields a *fundamental* cycle basis (FCB). Finding the minimum FCB is known to be **NP**-hard [13], but heuristics based on spanning trees prove to be very easy to implement and work reasonably well [13] (optionally, their cost can be improved by an edge-swapping phase [1, 31]).

5 The Eulerian cycle relaxation

In this section we construct a relaxation of Eq. (27). This is accomplished by substituting the $K|\mathcal{B}|$ cycle base constraints in Eq. (24) (occurring as the last line in Eq. (27)) with the K constraints obtained by considering a single Eulerian circuit in the given graph.

We follow a standard construction in order to find a Eulerian circuit, see e.g. [26]. We let G' be the multigraph obtained from G by adding sufficiently many parallel edges to G, so that the degree of each vertex in G' is even. This can always be done by [16], which implies that G' is Eulerian, i.e. it has a cycle incident with every edge in G' exactly once. We let $\mathcal{E}$ be a Eulerian cycle in G', and let $\overline{\mathcal{E}}$ be either of the two orientations of $\mathcal{E}$ obtained by walking over the cycle. We let $\overline{G}'$ be the digraph induced by the Eulerian circuit $\overline{\mathcal{E}}$. For each $\{i,j\} \in E$ let H_{ij} be the number of parallel edges between i, j in G'.

We note that $\overline{G}'$ might have parallel and antiparallel arcs. Consider the family of arc subsets $\mathcal{H}_{ij} = \{(i',j',h) \mid h \leq H_{ij} \wedge \{i',j'\} = \{i,j\}\}$ of $A(\overline{G}')$. We replace each arc $(i',j',h) \in \mathcal{H}_{ij}$ having $h > 1$ by an oriented 2-path $p_{i'j'h} = \{(i', v_{ijh}), (v_{ijh}, j')\}$ involving a new added vertex v_{ijh}. We call $\widetilde{G}$ the digraph obtained from $\overline{G}'$ with this replacement. We remark that $\widetilde{G}$ is simple (it has no parallel/antiparallel arcs) by construction. Moreover, $\widetilde{G}$ is a Eulerian digraph: take the Eulerian circuit $\overline{\mathcal{E}}$ in $\overline{G}'$, and, every time it traverses a parallel/antiparallel arc $(i',j',h) \in \mathcal{H}_{ij}$ with $h > 1$, let it traverse the oriented 2-path replacement $p_{i'j'h}$ instead: this is clearly a Eulerian circuit in $\widetilde{G}$, which we call $\mathcal{C}$.

Next we consider the simple graph $\widehat{G}$ obtained by replacing each arc in $\widetilde{G}$ with an (undirected) edge. Let $\widehat{V} = \{v_{ijh} \mid \{i,j\} \in E \wedge h > 1\}$, and $\widehat{E}$ be the subset of $E(\widehat{G})$ obtained by losing the orientation of the arcs in

$$\bigcup_{\substack{(i',j',h) \in \mathcal{H}_{ij} \\ \{i,j\} \in E \wedge h > 1}} p_{i'j'h}\,,$$

i.e., the union of all the edges from the 2-path replacements. We note that, by construction,

$$\widehat{V} = V(\widehat{G}) \smallsetminus V \quad \wedge \quad \widehat{E} = E(\widehat{G}) \smallsetminus E. \tag{28}$$

Let $c^{\mathcal{C}}_{ij} \in \{1, -1\}$ be the orientation of (i,j) in $\mathcal{C}$ w.r.t. $\widetilde{G}$; let $\widehat{\mathcal{C}}$ be the simple Eulerian cycle in $\widehat{G}$ corresponding to $\mathcal{C}$.

We can now prove the main result of this section.

▶ **Proposition 14.** *The formulation*

$$\left.\begin{array}{rrcl} \min_{s \geq 0, y} & \sum_{\{i,j\} \in E} (s^+_{ij} + s^-_{ij}) & & \\ \forall\, (i,j) \in A(\widetilde{G}) & \sum_{k \leq K} y^2_{ijk} - d^2_{ij} & = & s^+_{ij} - s^-_{ij} \\ \forall\, k \leq K & \sum_{(i,j) \in \mathcal{C}} c^{\mathcal{C}}_{ij} y_{ijk} & = & 0 \quad (\dagger) \end{array}\right\} \tag{29}$$

is a relaxation of Eq. (27)*.*

Proof. We first consider a variant of the cycle formulation in Eq. (27) applied to $\widehat{G}$, where, from the constraints corresponding to Eq. (8) (second line of Eq. (27)), we omit those indexed by $\widehat{E}$. We call this variant ($\star$). We *claim* that ($\star$) is an exact reformulation of Eq. (27) applied to G. The claim holds because $E(\widehat{G}) \smallsetminus \widehat{E} = E$ by Eq. (28), and because the signs of the y variables are irrelevant in Eq. (8) since they are squared. Now, since $\widehat{\mathcal{C}}$ is a Eulerian cycle in $\widehat{G}$, Eq. (†) must hold in $\widetilde{G}$ for any orientation of the edges of $\mathcal{C}$, by Lemma 1. Therefore, Eq. (†) is an aggregation of the constraints in Eq. (24), which occur within the reformulation ($\star$). So Eq. (29) is a relaxation of ($\star$). The proposition follows because of the *claim*. ◀

Note that Eq. (29) provides a solution $\overline{y}$ that may not satisfy Eq. (24), which also guarantee feasibility in Eq. (10) by Proposition 13. By Remark 12, this implies that Corollary 11 is no longer applicable. In other words, we cannot obtain a realization x of G from $\overline{y}$ using the linear system in Eq. (22), since $\overline{y}$ might well make Eq. (22) infeasibile. We can fix this issue by adjoining Eq. (22) to Eq. (29) as additional constraints. For practical reasons we also propose to adjoin the *centroid constraints*

$$\forall\, k \leq K \qquad \sum_{i \in V} x_{ik} = 0\,, \tag{30}$$

which provide a restriction of Eq. (27) by only keeping realizations of G having zero centroid (see Eq. (4)).

For a formulation P, we denote by $\mathsf{val}(P)$ its optimal objective function value.

▶ **Lemma 15.** *Let P be Eq.* (27)*, and P' be P with the x variables and the constraints in Eq.* (22) *adjoined. Then* $\mathsf{val}(P) = \mathsf{val}(P')$.

Proof. This is a direct consequence of Corollary 11. ◀

▶ **Lemma 16.** *For any reformulation (or relaxation) P of the EDGP involving the x variables, let P' be P with the centroid constraints Eq. (30) adjoined. Then* $\mathsf{val}(P) = \mathsf{val}(P')$.

Proof. Since P' is a restriction of P, and the optimization direction is minimization, we have $\mathsf{val}(P) \leq \mathsf{val}(P')$. Let x be an optimal solution of P: then $x' = x - \mathsf{stack}(\mathsf{centroid}(x), n)$ (where the second term of the right hand side is the centroid row K-vector stacked n times to yield an $n \times K$ matrix) is feasible in P' by definition, which proves that $\mathsf{val}(P) \geq \mathsf{val}(P')$. The result follows. ◀

We define the Eulerian cycle-based relaxation formulation, derived from Eq. (29) by adjoining Eq. (22) and Eq. (30), as follows:

$$\left.\begin{array}{rrcl} \min_{s\geq 0,x,y} & \sum_{\{i,j\}\in E}(s^+_{ij} + s^-_{ij}) & & \\ \forall\ (i,j) \in A(\widetilde{G}) & \sum_{k\leq K} y^2_{ijk} - d^2_{ij} & = & s^+_{ij} - s^-_{ij} \\ \forall\ k \leq K & \sum_{(i,j)\in \mathfrak{C}} c^{\mathfrak{C}}_{ij} y_{ijk} & = & 0 \\ \forall\ (i,j) \in A(\widetilde{G}) & x_{ik} - x_{jk} & = & y_{ijk} \\ \forall\ k \leq K & \sum_{i\in V} x_{ik} & = & 0. \end{array}\right\} \tag{31}$$

▶ **Proposition 17.** *Eq. (31) is a relaxation of the EDGP.*

Proof. Let us call Eq. (29) R and Eq. (27) P. By Proposition 14, R is a relaxation of P. By adjoining new variables x and Eq. (22) as constraints to both R and P, we obtain formulations R', P' such that R' is a relaxation of P'. But by Lemma 15 we have that $\mathsf{val}(P') = \mathsf{val}(P)$, so R' is a relaxation of P, which is a valid formulation of the EDGP. Note that Eq. (31) is R' with the centroid constraints Eq. (30) adjoined. By Lemma 16, therefore, $\mathsf{val}(R') = \mathsf{val}(31)$. Thus, Eq. (31) is a relaxation of the EDGP. ◀

▶ Remark 18. In general, we have $\mathsf{val}(31) \geq \mathsf{val}(29)$, since Lemma 15 only holds for Eq. (27), but not for Eq. (29), as mentioned under Proposition 14. Therefore Eq. (31) is a tighter relaxation than Eq. (29).

6 Computational experiments

The aim of this section is to compare the computational performance of the following EDGP formulations:

- **(i)** the cycle-based formulation in Eq. (27), where the realization is retrieved as a post-processing stage using (22) according to Corollary 11;
- **(ii)** the Eulerian cycle-based relaxation in Eq. (31);
- **(iii)** the classic edge-based formulation in Eq. (4).

All of these formulations are nonconvex Nonlinear Programs (NLP), which are generally NP-hard to solve. More specifically, all of these formulations are as hard to solve as the EDGP, which is NP-hard.

As a solution algorithm, we used a very simple MultiStart (MS) heuristic based on calling a local NLP solver from a random initial starting point at each iteration, and updating the best solution found so far as needed: although there are better heuristics around [12, 38, 46], MS is the best trade-off between implementation simplicity and efficiency. Moreover, more efficient heuristics often change the formulation during their execution, which may hinder the meaning of this computational comparison between formulations.

We evaluate the quality of a realization x of a graph G according to mean (MDE) and largest distance error (LDE), defined this way:

$$\mathsf{mde}(x, G) = \frac{1}{|E|} \sum_{\{i,j\}\in E} \big|\|x_i - x_j\|_2 - d_{ij}\big|$$
$$\mathsf{lde}(x, G) = \max_{\{i,j\}\in E} \big|\|x_i - x_j\|_2 - d_{ij}\big|.$$

Furthermore, for each realization x of a graph G found by using the MS algorithm, we consider the objective function value of the corresponding solution $\mathsf{solVal}(x, G)$. We note that, due to the heuristic nature of the MS, this value is not guaranteed to be globally optimal.

The CPU time taken to find the solution may also be important, depending on the application. In the control of underwater vehicles [3], for example, DGP instances might need to be solved in real time. In other applications, such as finding protein structure from distance data [8, 45] (our application of choice), the CPU time is not so important.

Our tests were carried out on a single CPU of a 2.1GHz 4-CPU 8-core-per-CPU machine with 64GB RAM running Linux. The local NLP solver used within the MS heuristic was the IPOpt solver [9]. We remarked in some preliminary tests that IPOpt was considerably slowed down by variants of Eq. (3) such as Eq. (5), which essentially move a nonconvexity on the objective to one in the constraints. The same holds for the cycle-based formulation in Eq. (27). We therefore reformulated Eq. (27) as follows:

$$\left.\begin{array}{rrcl} \min_y & \sum_{\{i,j\}\in A(\bar{G})}(\sum_{k\leq K} y_{ijk}^2 - d_{ij}^2)^2 & & \\ \forall\, k \leq K, B \in \mathcal{B} & \sum_{(i,j)\in A(B)} c_{ij}^B y_{ijk} & = & 0, \end{array}\right\} \tag{32}$$

and Eq. (31) similarly.

Our implementation consists of a mixture of Python 3 [49] and AMPL [18] interfaced through `amplpy`. Cycle bases and Eulerian cycles are found using `networkX` [21]. Solutions to the feasible but possibly overdetermined linear systems in Eq. (22) are obtained using an ℓ_1 error minimization approach reformulated as a Linear Programming problem solved with CPLEX [25].

6.1 Results

A benchmark on a diverse collection of randomly generated weighted graphs of small size and many different types, with a very similar set-up to the one discussed here, is presented in [34]. It was found that the cycle formulation finds better MDE values, while the edge formulation generally finds better LDE values and is faster. Some results on proteins, obtained with only 3 MS iterations, were also presented in [34].

The benchmark we consider here contains medium to large scale protein graph instances realized in $\mathbb{R}^3$, all of which contain cycles. W.r.t. the protein results presented in [34], we integrated one more instance, `1tii`, which, at 69800 edges and 5684 vertices, is considerably larger than all the others. The results are given in Tables 1 and 2.

In Table 1, we report instance name, instance sizes m and n, then performance measures MDE, LDE and CPU for cycle, Eulerian and edge-based formulations. In the last three lines we report average, standard deviation, and number of instances where the formulation performed best, for all performance measures. In all tested cases, finding the cycle basis, the Eulerian cycles, and solving Eq. (22) took a small fraction of the total solution time. The missing result for instance `100d` on the Eulerian cycle reformulation is due to a failure occurred in the `networkX` module because the graph of `100d` is not connected.

Table 1 Cycle formulation vs. Eulerian relaxation vs. edge formulation performances on protein graphs (realizations in $K = 3$ dimensions).

			MDE			LDE			CPU		
Instance	m	n	cycle	Eul	edge	cycle	Eul	edge	cycle	Eul	edge
`1guu`	955	150	0.086	0.069	**0.053**	1.234	1.068	**1.037**	**7.90**	553.76	290.21
`1guu-1`	959	150	0.080	0.082	**0.059**	1.013	1.069	**0.980**	9.67	23.03	**1.72**
`1guu-4000`	968	150	0.112	0.106	**0.092**	1.073	1.431	**0.936**	8.68	10.77	**1.56**
`pept`	999	107	**0.144**	0.239	0.179	2.862	**1.847**	1.943	5.52	4.72	**1.4**
`2kxa`	2711	177	**0.051**	0.119	0.172	3.705	**2.826**	3.813	21.53	25.54	**7.35**
`res_2kxa`	2627	177	**0.055**	0.237	0.156	**2.949**	3.570	3.054	20.84	21.20	**12.44**
`C0030pkl`	3247	198	**0.000**	0.145	0.211	**0.000**	3.537	3.829	29.50	26.69	**7.36**
`cassioli`	4871	281	0.146	0.113	**0.057**	3.914	3.616	**3.185**	47.23	48.44	**14.51**
`100d`	5741	488	**0.201**	-	0.251	**3.038**	-	3.987	387.32	-	**29.42**
`hlx_amb`	6265	392	**0.105**	0.214	0.119	3.836	3.888	**3.485**	120.25	80.27	**20.54**
`water`	11939	648	**0.146**	0.490	0.243	**3.579**	4.196	4.281	1346.69	399.42	**224.66**
`3al1`	17417	678	**0.062**	0.126	0.216	3.451	**3.175**	4.059	835.10	433.69	**123.45**
`1hpv`	18512	1629	**0.385**	0.402	0.416	3.847	**3.831**	4.015	10138.00	2387.29	**442.70**
`il2`	45251	2084	0.385	**0.049**	0.107	4.422	**4.204**	4.583	18141.22	9904.81	**5255.76**
`1tii`	69800	5684	0.620	0.436	**0.434**	6.755	4.492	**3.854**	18846.37	38230.21	**9039.28**
avg			**0.172**	0.202	0.184	**3.045**	3.054	3.136	3331.05	3724.99	**1031.49**
stdev			0.167	0.144	**0.118**	1.673	**1.204**	1.272	6672.49	10272.3	**2587.33**
|best|			**9**	1	5	4	5	**6**	1	0	**14**

It appears that, on average, there is relatively little difference between the quality performances of these three EDGP formulations on protein graphs of medium and large sizes. CPU-time wise, of course, the edge formulation is best. Cycle formulations, taken together, outperform the edge formulation on quality measures. The cycle-based formulation Eq. (27) is slightly better than the other formulations for both MDE and LDE. The number of instances on which Eq. (27) is best on quality measures is 13, against 11 for the edge-based formulation.

In Table 2, we report instance name and solVal for the cycle and the edge EGDP formulations. The three lines at the bottom of the table show the arithmetic and geometric mean of each column ("arithmean" and "geomean"), and the percentage of instances where the solution values of each formulation are smaller than the other ("best").

Table 2 MS solution values of cycle formulation vs. edge formulation (realizations in $K = 3$ dimensions).

	solVal	
Instance	cycle	edge
1guu	9.27E+02	**4.73E+02**
1guu-1	8.91E+02	**5.67E+02**
1guu-4000	1.40E+03	**1.01E+03**
pept	**3.21E+03**	3.50E+03
2kxa	**3.04E+03**	1.25E+04
res_2kxa	**3.42E+03**	9.81E+03
C0030pkl	**0**	1.92E+04
cassioli	2.37E+04	**7.73E+03**
100d	**3.16E+04**	4.36E+04
hlx_amb	2.04E+04	**1.97E+04**
water	**5.73E+04**	1.10E+05
3al1	**2.56E+04**	1.22E+05
1hpv	**2.71E+05**	3.03E+05
il2	7.76E+05	**1.46E+05**
1tii	2.33E+06	**1.23E+06**
arithmean	2.37E+05	1.36E+05
geomean	5.96E+02	1.86E+04
best	53.33%	46.67%

The cycle formulation reports better local optima more often than the edge formulation ("best" = 53.3%), while the latter is more stable, on average, as its arithmetic mean is slightly smaller. However, since the solution values of the cycle formulation are sometimes much smaller than those of the edge formulation, the geometric mean of the former is about two orders of magnitude smaller than that of the latter.

We observe that Eq. (27) was the only formulation by which a global optimum was found (that of C0030pkl) using MS. Overall, the results reported in Table 2 follow those of Table 1, except in the case of instance hlx_amb.

We decided to ignore the Eulerian formulation in Table 2, as its the objective function values were often larger than those of the corresponding cycle formulation, despite the fact that the former is a relaxation of the latter. This apparent anomaly is due to the heuristic nature of the MS solution algorithm.

All in all, we believe that our results show that cycle formulations are credible competitors w.r.t. the well established edge-based formulations, especially when the CPU time is not an important performance measure (which is generally the case in the protein conformation application).

References

1 E. Amaldi, L. Liberti, F. Maffioli, and N. Maculan. Edge-swapping algorithms for the minimum fundamental cycle basis problem. *Math. Methods Oper. Res.*, 69:205–223, 2009.

2 J. Aspnes, T. Eren, D. Goldenberg, S. Morse, W. Whiteley, R. Yang, B. Anderson, and P. Belhumeur. A theory of network localization. *IEEE Trans. Mobile Comput.*, 5(12):1663–1678, 2006.

3 A. Bahr, J. Leonard, and M. Fallon. Cooperative localization for autonomous underwater vehicles. *International Journal of Robotics Research*, 28(6):714–728, 2009.

4 N. Beeker, S. Gaubert, C. Glusa, and L. Liberti. Is the Distance Geometry Problem in NP? In A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, editors, *Distance Geometry: Theory, Methods, and Applications*, pages 85–94. Springer, 2013.

5 P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tightening techniques for non-convex MINLP. *Optim. Methods Softw.*, 24(4):597–634, 2009.

6 A. Berg and T. Jordán. Algorithms for graph rigidity and scene analysis. In G. Di Battista and U. Zwick, editors, *Algorithms: Proceedings of the European Symposium on Algorithms*, volume 2832 of *Lecture Notes in Computer Science*, pages 78–89. Springer, 2003.

7 M. Bukatin, R. Kopperman, S. Matthews, and H. Pajoohesh. Partial metric spaces. *Am. Math. Mon.*, 116(8):708–718, 2009.

8 A. Cassioli, B. Bordeaux, G. Bouvier, A. Mucherino, R. Alves, L. Liberti, M. Nilges, C. Lavor, and T. Malliavin. An algorithm to enumerate all possible protein conformations verifying a set of distance constraints. *BMC Bioinformatics*, 16:23–38, 2015.

9 COIN-OR. *Introduction to IPOPT: A tutorial for downloading, installing, and using IPOPT*, 2006.

10 R. Connelly. Generic Global Rigidity. *Discrete Comput. Geom.*, 33:549–563, 2005.

11 C. D'Ambrosio and L. Liberti. Distance Geometry in linearizable norms. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 10589 of *Lecture Notes in Computer Science*, pages 830–838. Springer, 2017.

12 C. D'Ambrosio, Ky Vu, C. Lavor, L. Liberti, and N. Maculan. New error measures and methods for realizing protein graphs from distance data. *Discrete Comput. Geom.*, 57(2):371–418, 2017.

13 N. Deo, G. M. Prabhu, and M. S. Krishnamoorthy. Algorithms for Generating Fundamental Cycles in a Graph. *ACM Trans. Math. Softw.*, 8(1):26–42, 1982.

14 G. Dias and L. Liberti. Diagonally dominant programming in distance geometry. In R. Cerulli, S. Fujishige, and R. Mahjoub, editors, *International Symposium in Combinatorial Optimization*, volume 9849 of *Lecture Notes in Computer Science*, pages 225–236. Springer, 2016.

15 Y. Ding, N. Krislock, J. Qian, and H. Wolkowicz. Sensor network localization, Euclidean distance matrix completions, and graph realization. *Optim. Eng.*, 11:45–66, 2010.

16 J. Edmonds and E. Johnson. Matching, Euler tours, and the Chinese postman. *Math. Program.*, 5:88–124, 1973.

17 T. Eren, D. Goldenberg, W. Whiteley, Y. Yang, A. Morse, B. Anderson, and P. Belhumeur. Rigidity, Computation, and Randomization in Network Localization. In *IEEE Annual Joint Conference: INFOCOM, IEEE Computer and Communications Societies*, pages 2673–2684. 2004.

18 R. Fourer and D. Gay. *The AMPL Book*. Duxbury Press, 2002.

19 A. Golynski and J. D. Horton. A polynomial time algorithm to find the minimum cycle basis of a regular matroid. In *8th Scandinavian Workshop on Algorithm Theory*, 2002.

20 S. Gortler, A. Healy, and D. Thurston. Characterizing generic global rigidity. *Am. J. Math.*, 132(4):897–939, 2010.

21 A. Hagberg, D. Schult, and P. Swart. Exploring network structure, dynamics, and function using `NetworkX`. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, 2008.

22 F. Harary. *Graph Theory*. Addison-Wesley Publishing Group, 1969.

23 P. Hoffman and B. Richter. Embedding graphs in surfaces. *J. Comb. Theory, Ser. B*, 36:65–84, 1984.

24 J. D. Horton. A Polynomial-Time Algorithm to Find the Shortest Cycle Basis of a Graph. *SIAM J. Comput.*, 16(2):358–366, 1987.

25 IBM. *ILOG CPLEX 12.9 User's Manual*. IBM, 2019.

26 D. Jungnickel. *Graphs, Networks and Algorithms*. Number 5 in Algorithms and Computation in Mathematics. Springer, 4 edition, 2013.

27 T. Kavitha, C. Liebchen, K. Mehlhorn, D. Michail, R. Rizzi, T. Ueckerdt, and K. Zweig. Cycle bases in graphs: characterization, algorithms, complexity, and applications. *Comput. Sci. Rev.*, 3:199–243, 2009.

28 M. Laurent. Cuts, matrix completions and graph rigidity. *Math. Program.*, 79:255–283, 1997.

29 C. Lavor, L. Liberti, and N. Maculan. Computational Experience with the Molecular Distance Geometry Problem. In J. Pintér, editor, *Global Optimization: Scientific and Engineering Case Studies*, pages 213–225. Springer, 2006.

30 C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Comput. Optim. Appl.*, 52:115–146, 2012.

31 J. Lee and L. Liberti. A matroid view of key theorems for edge-swapping algorithms. *Math. Methods Oper. Res.*, 76:125–127, 2012.

32 L. Liberti. A new distance geometry method for constructing word and sentence vectors. In *WWW'20: Companion Proceedings of the Web Conference 2020*, volume 20. ACM Press, 2020.

33 L. Liberti. Distance Geometry and Data Science. *Top*, 28:271–339, 220.

34 L. Liberti, G. Iommazzo, C. Lavor, and N. Maculan. A cycle-based formulation of the Distance Geometry Problem. In C. Gentile et al., editors, *Proceedings of 18th Cologne-Twente Workshop*, volume 4 of *AIRO Springer Series*.

Springer, 2020.

35 L. Liberti and C. Lavor. *Euclidean Distance Geometry: An Introduction*. Springer, 2017.

36 L. Liberti and C. Lavor. Open research areas in distance geometry. In A. Migalas and P. Pardalos, editors, *Open Problems in Optimization and Data Analysis*, volume 141 of *Springer Optimization and Its Applications*, pages 183–223. Springer, 2018.

37 L. Liberti, C. Lavor, J. Alencar, and G. Abud. Counting the number of solutions of ^{K}DMDGP instances. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 224–230. Springer, 2013.

38 L. Liberti, C. Lavor, N. Maculan, and F. Marinelli. Double Variable Neighbourhood Search with smoothing for the Molecular Distance Geometry Problem. *J. Glob. Optim.*, 43:207–218, 2009.

39 L. Liberti, C. Lavor, N. Maculan, and A. Mucherino. Euclidean distance geometry and applications. *SIAM Rev.*, 56(1):3–69, 2014.

40 L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *Int. Trans. Oper. Res.*, 18:33–51, 2010.

41 L. Liberti, N. Mladenović, and G. Nannicini. A good recipe for solving MINLPs. In V. Maniezzo, T. Stützle, and S. Voß, editors, *Hybridizing metaheuristics and mathematical programming*, volume 10 of *Annals of Information Systems*, pages 231–244. Springer, 2009.

42 L. Liberti and K. Vu. Barvinok's naive algorithm in distance geometry. *Oper. Res. Lett.*, 46:476–481, 2018.

43 C. Liebchen and R. Rizzi. A greedy approach to compute a minimum cycle basis of a directed graph. *Inf. Process. Lett.*, 94:107–112, 2005.

44 L. Lovász and M. Plummer. On minimal elementary bipartite graphs. *J. Comb. Theory, Ser. B*, 23:127–138, 1977.

45 T. Malliavin, A. Mucherino, C. Lavor, and L. Liberti. Systematic exploration of protein conformational space using a distance geometry approach. *J. Chem. Infor. Mod.*, 59:4486–4503, 2019.

46 L. Mencarelli, Y. Sahraoui, and L. Liberti. A multiplicative weights update algorithm for MINLP. *EURO J. Comput. Optim.*, 5:31–86, 2017.

47 A. Mucherino, C. Lavor, and L. Liberti. Exploiting symmetry properties of the discretizable molecular distance geometry problem. *J. Bioinfor. Comput. Biol.*, 10, 2012.

48 K. Paton. An Algorithm for Finding a Fundamental Set of Cycles of a Graph. *Commun. ACM*, 12(9):514–518, 1969.

49 G. van Rossum et al. *Python Language Reference, version 3*. Python Software Foundation, 2019.

50 J. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. In *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489. 1979.

51 G. Schaeffer. Random sampling of large planar maps and convex polyhedra. In *STOC'99: Proceedings of the thirty-first annual ACM symposium on Theory of Computing*, pages 760–769. ACM Press, 1999.

52 S. Seshu and M. B. Reed. *Linear Graphs and Electrical Networks*. Addison-Wesley Publishing Group, 1961.

53 A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30:20–36, 2011.

54 M. Tawarmalani and N. V. Sahinidis. Global Optimization of Mixed Integer Nonlinear Programs: A Theoretical and Computational Study. *Math. Program.*, 99:563–591, 2004.