Reconstructing a Three-Dimensional Model with Arbitrary Errors

Bonnie Berger* **Jon Kleinberg†** **Tom Leighton‡**

Abstract

A number of current technologies allow for the determination of inter-atomic distance information in structures such as proteins and RNA. Thus, the reconstruction of a three-dimensional set of points using information about its inter-point distances has become a task of basic importance in determining molecular structure. The distance measurements one obtains from techniques such as NMR are typically sparse and error-prone, greatly complicating the reconstruction task. Many of these errors result in distance measurements that can be safely assumed to lie within certain fixed tolerances. But a number of sources of systematic error in these experiments lead to inaccuracies in the data that are very hard to quantify; in effect, one must treat certain entries of the measured distance matrix as being arbitrarily "corrupted."

The existence of arbitrary errors leads to an interesting sort of error–correction problem — how many corrupted entries in a distance matrix can be efficiently corrected to produce a consistent three-dimensional structure? For the case of an $n \times n$ matrix in which every entry is specified, we provide a randomized algorithm running in time $O(n \log n)$ that enumerates all structures consistent with at most $(\frac{1}{2} - \varepsilon)n$ errors per row, with high probability. In the case of randomly located errors, we can correct errors of the same density in a sparse matrix — one in which only a β fraction of the entries in each row are given, for any constant $\beta > 0$.

1 Introduction

The reconstruction of a three–dimensional set of points using information about its inter-point distances is a task of basic importance in determining molecular structure. In particular, spectroscopic methods such as two–dimensional NMR provide a means for determining a labeled subset of the distances between atoms in large structures such as proteins and RNA, and such methods have therefore proved extremely valuable in conformational studies of such molecules (see e.g. [19, 22]).

*Mathematics Department and Lab. for Computer Science, MIT. Email: bab@theory.lcs.mit.edu.

†This work was done while at Laboratory for Computer Science, MIT, Cambridge, MA 02139. Present address: Department of Computer Science, Cornell University, Ithaca NY 14853. Email: kleinber@cs.cornell.edu.

‡Mathematics Department and Lab. for Computer Science, MIT. Email: ftl@math.mit.edu.

Determining the structure of n labeled points in $\mathbf{R}^3$ is easy if one is given the exact value of every inter-point distance. However, this is far from what one obtains using NMR and its variants. In particular, not all distances can be measured, and the distance values that one does obtain are susceptible to a wide variety of sources of error, not all of which are fully understood [12, 14, 17, 20, 23]. Below, we discuss some of the ways in which errors arise in such experiments. For now, we simply emphasize that many of these errors are not "statistical" in nature; rather, they are systematic errors arising from fundamental limitations in the underlying technique. Moreover, while some of these errors produce distances that can be safely assumed to lie within known tolerances, others appear to result in essentially arbitrary and unpredictable numerical values.

Thus, at its most general level, the problem we are dealing with is the following. We are given an $n \times n$ symmetric matrix M for which, in principle, there is some set P of n labeled points in $\mathbf{R}^3$ so that $M_{ij} = d(p_i, p_j)$. (Here M_{ij} denotes the $(i,j)^{\text{th}}$ entry of M, and $p_i, p_j \in P$.) The goal is to determine P, up to congruence. Our task is complicated by the following facts.

(i) The matrix M is incomplete (i.e. some entries are "blank.")

(ii) Most of the entries of M are only known to within certain fixed tolerances.

(iii) A limited number of the entries may be arbitrarily "corrupted," and bear essentially no relation to the corresponding inter-point distance.

Problems arising from (i) alone — exact but sparse distance information — forms the core of *rigidity theory*, a subject dating back to mathematicians and mechanical engineers of the nineteenth century (e.g. [3]). Problems arising from (i) and (ii) together are currently handled in practice by algorithms based on *distance geometry* (see e.g. [6, 7]), which has proved to be effective at refining an initial set of distance measurements by a variety of constraint propagation and "bounds smoothing" techniques.

In this paper we consider the third issue, that of correcting arbitrary errors in distance measurements so as to correctly recover a three-dimensional structure. In addition to the fact that it arises naturally in the analysis of NMR data, this is a very basic algorithmic problem in geometry. In effect, we are asking for an error-correcting algorithm in this geometric setting; given a corrupted distance matrix, how efficiently can one "decode" it to a correct point set in three-space? The crucial point here is that we do not *know* the location of the errors, and we want an algorithm that is efficient even when the number of errors is quite large.

We develop theoretical tools for addressing questions of this type, and provide two error-correcting algorithms for distance matrices. We provide a randomized algorithm with running time $O(n \log n)$ that corrects up to $(\frac{1}{2} - \varepsilon)n$ corrupted entries per row in a distance matrix for any constant $\epsilon > 0$. We then extend the algorithm to the case of arbitrarily sparse matrices: we can correct randomly located corruptions of the same density in a matrix in which only βn entries per row are given, for arbitrarily small $\beta > 0$. Thus, we are able to correct distance matrices in which up to 49% of the entries in every row are incorrect; in a sense that we describe below, this error density is essentially optimal. The algorithms themselves are quite natural; their analysis involves techniques based on random sampling, rigidity theory, and some basic extremal graph theory.

We emphasize at the outset that the model in which we are working is too idealized to be directly applicable in the NMR applications discussed above. Specifically, we deal with matrices in which every entry is either correct or arbitrarily corrupted, and thus are not directly incorporating the notion of error tolerances for the "uncorrupted" entries. However, we feel that this abstraction makes sense for the basic goal of understanding the structure of these error–correcting problems; moreover, the algorithms and their analysis seem robust enough to suggest approaches to the other problems as they arise in applications. In particular, our algorithms do not appear to rely very heavily on having exact data; and we have recently been looking at some heuristics to incorporate the notion of error thresholds into our basic approach. In this regard, it also seems promising to consider the use of our approach together in some fashion with the distance geometry techniques of [6, 7], as the two problems they address are essentially orthogonal. We will have more to say about this in Section 1.4.

1.1 Our Results

We first consider the most basic problem — every entry of M is specified, and all but at most k entries are correct. For how large a value of k does there still exist a unique point set P whose distance matrix differs from M in at most k entries? And for how large a value of k can we find this P in polynomial time? We show

(1) *Two correct distance matrices that are not equal differ in at least $n-4$ entries. This bound is tight in the sense that there exist pairs of correct matrices that differ in exactly $n-4$ entries.*

(2) *Let $\varepsilon > 0$. Given an $n \times n$ matrix that differs from a correct distance matrix in at most $(\frac{1}{2}-\varepsilon)n$ corrupted entries, we can produce the point set (unique up to congruence) that generates this correct distance matrix in $O(n)$ time with arbitrarily high probability.*

Note that in linear time, one can only read a vanishingly small fraction of the input matrix. Also, note that (2) is tight in the sense that a matrix could differ from two non–congruent point sets in $\frac{1}{2}(n-4)$ entries. (This follows from (1).)

If one looks at the pairs of correct distance matrices that differ in very few entries, one finds that these "disagreements" tend to be clustered in a very small number of rows. This suggests that one should be able to correct many more errors if one assumes that at most a fixed fraction of the entries in any row of M are corrupted; and this turns out to be the case. We show

(3) *Let $\varepsilon > 0$. Given an $n \times n$ matrix M, there are at most a constant number of correct distance matrices that differ from M in at most $(\frac{1}{2}-\varepsilon)n$ entries* per row, *and point sets corresponding to all these correct matrices can be enumerated in $O(n \log n)$ time with arbitrarily high probability.*

Algorithmically, this is a much more powerful result than (2) — even if 49% of the entries in every row are arbitrarily corrupted, one can still recover all consistent three–dimensional structures in $O(n \log n)$ time. Moreover, this result is tight in the sense that there are matrices M for which there are infinitely many correct distance matrices differing from M in $\frac{1}{2}n$ entries per row.

Finally, we can extend this algorithm to handle arbitrarily sparse matrices — those in which βn entries per row are given, for some small $\beta > 0$. Here we consider a symmetric matrix M, and build a graph G_M on the rows (or columns) of M, joining rows i and j if M_{ij} is given as an entry (whether or not it is correct).

Handling sparsity poses a number of additional problems, and thus our algorithm here will not be quite as general as it was in the case of complete data. First, although we allow the structure of G_M to be arbitrary, we need to assume random location of errors: each given entry is corrupted with probability $(\frac{1}{2}-\varepsilon)$. Also note that because of the sparsity assumption, it is possible that G_M is disconnected, in which case there is no hope of finding a finite number of candidate point sets. Hence it is necessary to make some sort of "connectivity" restriction; we require that G_M have constant vertex-connectivity, for a constant depending on the sparsity.

Given this, we show

(4) *For all $\beta, \varepsilon, \delta > 0$, there is a constant κ such that the following holds. Let M be a correct distance matrix with at least βn non-blank entries per row, and for which the graph G_M defined above is κ-vertex-connected. Let M' be a matrix obtained by corrupting each non-blank entry with probability $(\frac{1}{2} - \varepsilon)$ to an* arbitrary *value. Then a constant number of correct distance matrices that includes M can be enumerated by a randomized algorithm with probability at least $1 - \delta$ in $O(n \log n)$ time.*

1.2 Errors in NMR Data

In this section we give a brief overview of some basic NMR techniques, both for the sake of background and to give a sense for some of the sources of error in NMR. More detailed descriptions can be found in any of a number of sources; see e.g. [19, 22]. NMR experiments are frequently based on measuring proton resonances, since hydrogen atoms are easily detected by these techniques; this is in general sufficient for conformational studies, since hydrogen atoms are abundant in organic molecules. To determine spatial information, one typically requires two NMR experiments: one based on spin-spin coupling to determine protons that are close in the covalent geometry of the molecule (e.g. COSY, *Correlated Spectroscopy*), and another to determine distance information for pairs of protons that are close in space but not necessarily close in the covalent structure (e.g. NOESY, *Nuclear Overhauser Enhancement Spectroscopy*). Combining information from these two types of experiments, one obtains *labeled* inter-atomic distance data (i.e. one determines which distance corresponds to which pair of atoms).

The *nuclear Overhauser effect* (NOE), on which the latter technique is based, is manifested in cross-peaks in an NMR spectrum that arise from dipole-dipole coupling of one proton with nearby protons. NOE intensity is given by a formula with a term for d_{ij}^{-6}, where d_{ij} is the Euclidean distance between two protons; if one has approximate knowledge of the other variables involved, then it is feasible to recover a value for d_{ij}.

Distance calculations based on NOESY are subject to numerous sources of systematic error. Some basic examples are the following.

(i). Spin–diffusion is a phenomenon whereby NOE effects between protons i and j are in some sense "transferred" through a third proton k, resulting in spurious cross-peaks. Discussing the problem of structure determination of bound ligands, [14] writes, "Spin–diffusion pathways can also give rise to NOE cross-peaks that cannot be assigned to direct distances between bound-ligand protons. These effects often result in a compression of the apparent distances derived from transferred NOE experiments."

(ii). For reasons of experimental efficiency, the relaxation delays between successive NMR scans are typically too short to allow for full recovery of the magnetization effects between nearby protons [12, 17, 23]. This is particularly a problem in RNA structure determination, where proton relaxation times are quite long and vary considerably [12].

(iii). As the NMR spectrum becomes more and more dense, it is possible for two peaks to essentially coincide, and thus for one or more of these peaks to escape detection when the data is interpreted [20]. This is a problem that becomes increasingly significant precisely when more of the interpoint distances are measurable.

(iv). As noted above, in order to produce *labeled* distance data from a NOESY experiment — i.e., to determine that a given distance d is in fact associated with protons i and j — one must solve a complex assignment problem using information from an accompanying COSY experiment [19]. This is by no means a straightforward process; and clearly, an error in the assignment process will lead to inaccuracies in the distance matrix that are very hard to quantify.

In several of these cases, therefore, it may make sense to assume that certain of the final distance measurements one obtains are incorrect in a completely "arbitrary" fashion. Again, we do not suggest that the algorithms developed in this paper can be applied directly to deal with the issues discussed in this section. But it seems promising to explore how our basic approach could be used to handle a small number of egregious errors in distance data, together with lower and upper bounds on the remaining, more accurate, data.

1.3 Macroscopic Applications

Reconstructing sets of points in three dimensions from faulty data also arises in macroscopic applications. For example, the National Geodetic Survey (NGS) periodically is faced with the monumental task of updating the North American Datum, a network of reference points for which it is necessary to know exact latitude, longitude and altitude. Unfortunately, many exisiting measurements are in error and these errors propagate over time if new points are added that are located based on old faulty data. (Such problems also arise in surveying.) The traditional approach to this problem [11, 10] is to attempt solve a system of millions of nonlinear equations in order to find locations for the points that minimize an error function. The approach in this paper is quite different but tries to achieve the same end result with a much simpler algorithm. Similar issues arise in the analysis of satellite ranging data [21].

1.4 Related Geometric Algorithms

Algorithms developed by Havel and others in the area of *distance geometry* (e.g. [6, 7]) are used in practice for "refining" distance measurements that are initially known to lie between given lower and upper bounds (ℓ_{ij}, u_{ij}). The algorithm described in [7], for example, determines the tightest set of constraints consistent with the initial bounds and the triangle inequality; it also performs further constraint propagation using the four–point *tetrangle inequality*. Determining the tightest set of constraints consistent with Euclidean embeddability remains open.

This approach naturally handles sparse data, by setting the lower–upper bound pair for a blank entry to be $(0, \infty)$. However, the algorithm of [7] does not provide a systematic method for correcting arbitrary errors of the type we consider here. Thus, we feel that an interesting direction for future work would be to investigate effective ways of combining the approach here with the techniques of distance geometry.

Within the area of rigidity theory — when a sparse, error–free, set of distances is known — the work most closely associated with problems of reconstruction is that dealing with *global rigidity* [4, 9, 10]. Here, the issue is to determine when a set of exact distances is sufficient to uniquely determine a point set up to congruence. Finally, we note related work of Saxe [15], proving that it is NP-hard to decide whether a sufficiently sparse set of distances is consistent with a set of points in $\mathbf{R}^d$. This contrasts with the results of Section 2, which imply that if the set of distances is sufficiently dense, then the set of all consistent conformations can be enumerated efficiently.

2 Preliminary Results: The Fault–Free Case

For $x, y \in \mathbf{R}^3$, let $d(x, y)$ denote the standard Euclidean distance between them. If P is a set of n labeled points $\{p_1, \ldots, p_n\}$ in $\mathbf{R}^3$, we use $D(P)$ to denote the symmetric $n \times n$ matrix whose $(i, j)^{\text{th}}$ entry is the distance $d(p_i, p_j)$. For two labeled n-point subsets P and Q of $\mathbf{R}^3$, we will say that P is *congruent* to Q, written $P \sim Q$, if Q can be obtained from P by a rigid motion, or by a rigid motion composed with a reflection. By a *conformation*, we mean an equivalence class of point sets under congruence; by abuse of notation, we will sometimes use a fixed point set interchangeably with its conformation.

Throughout the remainder of the paper, we frequently make the following two non-degeneracy assumptions about any point set P in $\mathbf{R}^3$:

(1) No four points of P lie on a plane.

(2) No ten points of P lie on a quadric surface.

These assumptions are quite common in computational geometry, and the analysis to follow will make use of them in places.

Here are some basic facts that we will use in what follows. In all these lemmas, let P and Q be non-degenerate sets of n points in $\mathbf{R}^3$.

Lemma 2.1 *Suppose that* $|P| = |Q| \geq 4$, $P \sim Q$, *and let* $x, y \in \mathbf{R}^3$ *be such that* $d(x, p_i) = d(y, q_i)$ *for* $i = 1, \ldots, 4$. *Then* $P \cup \{x\} \sim Q \cup \{y\}$.

Lemma 2.2 *If* $D(P) = D(Q)$, *then* $P \sim Q$.

Of course, the converse of Lemma 2.2 also holds.

We begin by analyzing rigidity properties of point sets in the special case when the distance matrix is incomplete but contains no errors. These preliminary results will prove useful when we later consider the case with faulty data. For the time being, we will not worry about algorithmic issues, since all of the results presented in this section can be made to run in near linear time using the algorithms that will be developed in subsequent sections.

In what follows, we let M denote a symmetric matrix in which some entries are blank. We let $G_M = (V, E)$ denote the graph on the n points that has an edge (i, j) if the entry M_{ij} is not blank; each edge is labeled with this corresponding value. In this section, we will write G to denote G_M. We say that a conformation P is *consistent with* M (or with G_M) if for each pair (i, j) with M_{ij} non-blank, we have $d(p_i, p_j) = M_{ij}$. Let d denote the minimum degree of the nodes in G.

2.1 The Case When $d \geq \frac{3n}{4} + 1$

In the case that G is sufficiently dense, there can only be one conformation consistent with M.

Theorem 2.3 *If $d \geq \frac{3n}{4} + 1$, then there is a unique conformation that agrees with M.*

Proof. First observe that we may assume $n > 8$, or else the result simply follows from Lemma 2.2. We will show that there exists an ordering $v_1, v_2, \ldots, v_n$ of the nodes of G so that $\{v_1, v_2, v_3, v_4\}$ are all adjacent in G and so that for all $i \geq 4$, v_{i+1} is adjacent to at least 4 nodes in $\{v_1, v_2, \ldots, v_i\}$. The theorem then follows by repeated applications of Lemma 2.1.

We start by selecting an arbitrary vertex v_1. Then, choose v_2 to be any of the (at least) $\frac{3n}{4} + 1$ neighbors of v_1, choose v_3 to be any of the (at least) $\frac{n}{2} + 2$ neighbors of both v_1 and v_2, and choose v_4 to be any of the (at least) $\frac{n}{4} + 3$ neighbors of v_1, v_2, and v_3.

We next proceed to find any node v_{i+1} that has at least 4 edges to $\{v_1, v_2, \ldots, v_i\}$, iteratively for all $i < n$. To show that such a node always exists, we note that the number of edges leaving the set $S_i = \{v_1, v_2, \ldots, v_i\}$ is at least $(\frac{3n}{4} + 1)i - (i - 1)i = (\frac{3n}{4} - i + 2)i$.

For $4 \leq i \leq \frac{1}{2}n$, this is greater than $3(n - i)$, and hence when i is in this range we can find such a v_{i+1} with 4 neighbors in the set S_i. And when $i > \frac{1}{2}n$, any point not already in the set S_i can be used as v_{i+1}, since any such point has at least

$$\frac{3n}{4} + 1 - \left(\frac{n}{2} - 1\right) = \frac{n}{4} + 2 \geq 4$$

neighbors in S_i. (The last inequality follows from our assumption that we may consider the case $n > 8$.) ∎

Corollary 2.4 *If the distance matrices for two point sets P and Q agree in at least $\frac{3n}{4} + 1$ entries per row, then $P \sim Q$.*

2.2 The Case When $d \le \frac{3n}{4}$

When $d \le \frac{n}{2} - 1$, there can be an infinite number of distinct conformations consistent with a distance matrix M, as is evidenced by the example in which G is the union of two disjoint copies of $K_{n/2}$. Thus, it is natural to ask what can be said about the range $\frac{n}{2} \le d \le \frac{3n}{4}$. Here, a result of Connelly (unpublished, [5]) implies that the graph $K_{6,6}$ is generically globally rigid in $\mathbf{R}^3$; that is, any non–degenerate labeling of distances on its edges results in a unique consistent conformation. Also, a result from extremal graph theory (see the proof of Theorem 2.5) implies that if n is large enough and $d \ge \frac{n}{2} + 4$, then G contains a copy of $K_{6,6}$. Combining these two facts with the method used to prove Theorem 2.3, one can show that in fact when $d \ge \frac{n}{2} + 4$, there is a unique conformation consistent with any (non–degenerate) distance matrix M.

Even when d is $o(n)$, however, it is still possible to characterize the possible conformations, as we show in the following theorem.

Theorem 2.5 *If $d \ge \alpha n^{4/5}$ for some sufficiently large constant α, then G can be partitioned into $O(n/d)$ subgraphs, each of size $\Omega(d)$ and each having $O(1)$ consistent conformations.*

Proof. We make use of a theorem of Bolker and Roth [1] stating that, subject to our non-degeneracy assumptions (1) and (2), the graph $K_{4,6}$ is rigid in $\mathbf{R}^3$ for any set of distances on its edges. (See also [18].) By standard results of Milnor [13] and Thom [16], this implies that a given distance-labeled $K_{4,6}$ has $O(1)$ consistent conformations.

A well-known result in extremal graph theory (see Bollobas [2, ch. 6]) states that if the number of edges of a graph exceeds $s^{1/t}n^{2-1/t} + tn$, then the graph contains a copy of $K_{s,t}$, for any s and t. Thus, we know that G has a sufficient number of edges so that it contains a subgraph isomorphic to $K_{4,8n/d}$. This subgraph can be used as a seed to grow a rigid component G_1 with at least $d/2$ points by the repeated addition of points that have at least 4 neighbors in G_1. (The proof that G_1 can be grown to size at least $d/2$ consists of a simple pigeonhole principle argument much like that used to prove Theorem 2.3.) Once G_1 can be grown no further, it is removed from G and we begin again. The key to continuing the process is to note that the minimum degree of nodes in $G - G_1$ is $d - 3$ since (by definition of G_1) each node in $G - G_1$ has at most 3 neighbors in G_1.

Continuing in this fashion, we cover the nodes of G with $O(n/d)$ disjoint subgraphs $G_1, G_2, \ldots, G_{O(n/d)}$, each of size $\Omega(d)$. By the argument of the first paragraph (applied to the initial $4 + 8n/d$ nodes), combined with Lemma 2.1, we see that each G_i has at most $O(1)$ consistent conformations. ■

Since it is possible that the G_i may not be connected to one another, there is nothing more of a general nature that can be said about the conformations of G. If, however, G is highly connected, then it is possible to pin down (at least to a certain extent) the conformations of each subgraph with respect to the others. This technique will be used in Section 4.

3 Full Distance Matrices with Errors

We next analyze the scenario when distances are present in M for all pairs of points but when many of the entries are incorrect. We will say that a conformation is γ-*consistent with*

M if its distance matrix agrees with M in at least a γ fraction of the entries per row. Our main result is a randomized algorithm for producing all distinct conformations of points that are γ-consistent with M, for any $\gamma \geq \frac{1}{2} + \varepsilon$. The algorithm runs in $O(n \log n)$ time and with probability arbitrarily close to 1 outputs a set of conformations containing all those that are γ-consistent with M.

As a consequence of the analysis of the algorithm, we will find that there are at most $O(1)$ conformations that are γ-consistent with any full distance matrix, for any $\gamma \geq \frac{1}{2} + \varepsilon$. This result is tight in two respects. First, there is a matrix with an infinite number of distinct $\frac{1}{2}$-consistent conformations. For example, consider a distance matrix where the distances between points $p_1, p_2, \ldots, p_{n/2}$ and $p_{n/2+1}, \ldots, p_n$ are incorrect. In this case, there are an infinite number of ways that $p_1, p_2, \ldots, p_{n/2}$ can be arranged in relation to $p_{n/2+1}, \ldots, p_n$.

Second, there is a matrix with two distinct $\frac{3}{4}$-consistent conformations. For example, consider two non-congruent conformations P and Q for which the pairwise distances among $p_1, p_2, \ldots, p_{n/2}$ equal those among $q_1, q_2, \ldots, q_{n/2}$ and for which the pairwise distances among $p_{n/2+1}, \ldots, p_n$ equal those among $q_{n/2+1}, \ldots, q_n$. Let M be the matrix that correctly reports the distances among $p_1, p_2, \ldots, p_{n/2}$ and among $p_{n/2+1}, \ldots, p_n$. For the other entries of M, choose half of the values per row to be consistent with P and half to be consistent with Q.

Corollary 2.4 implies directly that when at least a $\gamma' = (\frac{7}{8} + \varepsilon)$ fraction of the entries of each row are correct, there is a unique γ'-consistent conformation. And from the generic global rigidity of $K_{6,6}$ (see the remark at the beginning of Section 2.2), it is possible to show that when there is at least a $\gamma'' = (\frac{3}{4} + \varepsilon)$ fraction of correct entries in each row, there is a unique γ''-consistent conformation. This is tight, given the example of the previous paragraph.

In any event, the algorithm will find all consistent conformations with high probability.

3.1 The Algorithm

In what follows, we let P denote any point set that is $(\frac{1}{2} + \varepsilon)$-consistent with with M, and in the remainder of this section the term "consistent" will (by abuse of notation) stand for "$(\frac{1}{2} + \varepsilon)$-consistent." We use H to denote the graph that indicates which distances in M match those in P; that is, two points are connected by an edge on H if their distance in M matches their distance in P. We do not know P or H, of course, but we do know that every point in H has degree at least $(\frac{1}{2} + \varepsilon)n$. Also, there might be several different H's, each corresponding to a different (but consistent) point set P. In what follows, we will describe a randomized algorithm that outputs $O(1)$ consistent point sets (including P) with constant probability. Since the result holds for any P and does not use knowledge of P, this will mean that there are at most $O(1)$ possible non-congruent choices for P.

The algorithm itself is quite simple. We start by randomly ordering the points $\{v_1, v_2, \ldots, v_n\}$. Let c be a constant, and consider the following two events.

(i) the subgraph induced on $\{v_1, v_2, \ldots, v_c\}$ has minimum degree $d_c \geq \frac{1}{2}c + 4$.

(ii) for all $i \geq c$, v_{i+1} is connected to at least $(\frac{1}{2} + \frac{1}{2}\varepsilon)i$ nodes in $\{v_1, v_2, \ldots, v_i\}$.

For a constant $\lambda_0 < 1$, the probability that the first event fails to happen is at most $c\lambda_0^c$, and the probability that the second fails to happen is at most $\sum_{i \geq c} \lambda_0^i$. Thus by choosing c large enough, the probability that both events happen can be made arbitrarily close to 1.

If the ordering satisfies conditions (i) and (ii) above, then we say that the ordering is *good.* Of course, we do not know whether or not any ordering is good, but we do know that almost all orderings are good by the definition of c. In what follows, we will show how to construct a set of $O(1)$ consistent conformations that includes P with very high probability given any good ordering.

We begin by producing all the conformations for $\{v_1, v_2, \ldots, v_c\}$ that are consistent with $M_c = \{M_{ij} \mid 1 \leq i, j \leq c\}$ assuming that there are at most $\frac{1}{2}c - 4$ errors per row in M_c. These conformations are produced as follows.

Let H_c denote the graph of correct distances in M_c. Since $d_c \geq \frac{1}{2}c + 4$, we know by the result from [2] (see the proof of Theorem 2.5 above), that H_c contains a copy of $K_{4,6}$ (provided that c is a sufficiently large constant). As before, this graph will be infinitesimally rigid and hence have at most $O(1)$ conformations [1, 13, 16]. One of these conformations will agree with P.

Next, we order the vertices of H_c so that $\{v_1, v_2, \ldots, v_{10}\}$ form a copy of the $K_{4,6}$, and so that for all $10 \leq i < c$, v_{i+1} has at least 4 connections to $\{v_1, v_2, \ldots, v_i\}$. The argument that such an ordering exists is strictly analogous to the proof of Theorem 2.3, using the fact that $d_c \geq \frac{1}{2}c + 4$. As a consequence of Lemma 2.1, we can conclude that H_c is rigid and has $O(1)$ conformations, one of which matches P. Unfortunately, we do not know what the edges of H_c are and so we must try all possibilities. This takes at most $O(1)$ time, however, since c is a constant. As a result of this process, we produce a large constant number of potential conformations, including one that agrees with P on $\{v_1, \ldots, v_c\}$.

We next show how to extend each of the potential conformations in a unique way to the entire n-point set, so that the conformation that agrees with P on $\{v_1, v_2, \ldots, v_c\}$ will extend to P. We will make heavy use of condition (ii) above, even though we do not know which distances from each v_{i+1} to $\{v_1, v_2, \ldots, v_i\}$ are the correct distances (i.e., those that match P).

Once an initial conformation has been extended to locate $\{v_1, v_2, \ldots, v_i\}$ for $i \geq c$, we will locate v_{i+1} as follows. For simplicity of exposition, let us first consider the case in which i is a sufficiently large multiple of $\log n$. We first select $\alpha \log n$ 4-tuples from $\{v_1, v_2, \ldots, v_i\}$ at random, where α is a sufficiently large constant. Each 4-tuple is used to create a candidate location for v_{i+1}. If the conformation of $\{v_1, v_2, \ldots, v_i\}$ agrees with P, then with probability $1 - O(1/n^2)$, nearly 1/16 of the candidate locations for v_{i+1} will coincide at a single point matching that of the location of v_{i+1} in P. At most $O(1)$ candidate locations can have been produced nearly 1/16 of the time and they can be identified in $O(\log n)$ steps.

The correct location for v_{i+1} can then be determined in $O(\log n)$ additional steps as follows. The distance is computed between each of the $O(1)$ remaining candidates and $\alpha \log n$ randomly selected points $R \subset \{v_1, v_2, \ldots, v_i\}$. These distances are then compared with the corresponding values in M. We then select the unique candidate position that matches at least $\frac{1}{2}|R| + 4$ of the distances in M. There exists such a candidate with high probability, since with probability $1 - O(1/n^2)$ at least a $(\frac{1}{2} + \frac{1}{4}\varepsilon)$ fraction of the distances selected from M for each candidate are correct (and one of the candidates is correct); and there cannot be two such candidates since any two such candidates must share 4 common correct distances, which means that they must have the same location. If any of the preceding conditions is violated, then we can terminate further consideration of the underlying conformation since we know with high probability that it does not agree with P. Finally, if $i < \alpha \log n$, then the probability of failure is at most λ_1^i, for a constant λ_1; thus the probability of failure for any i between c and $\alpha \log n$ is at most $\sum_{i \geq c} \lambda_1^i$, which is arbitrarily close to 0.

At the end of $O(n \log n)$ steps, the preceding algorithm will have produced a set of $O(1)$ potential conformations which includes P with probability close to 1. We next enter a phase where we check to make sure that each target conformation is at least $(\frac{1}{2} + \frac{1}{2}\varepsilon)$-consistent with M. The check is done by random sampling. Specifically, for each v_i, we choose $\alpha \log n$ other points at random (where α is a sufficiently large constant) and check that a $(\frac{1}{2} + \frac{3}{4}\varepsilon)$ fraction of the distances to these points match the corresponding values in M. If the match is successful for every v_i, then we can be assured with probability $1 - O(1/n^2)$ that the conformation agrees with M on at least a $(\frac{1}{2} + \frac{1}{2}\varepsilon)$ fraction of the entries in each row. Similarly, P will be found to be consistent with M with high probability.

In conclusion, we have shown

Theorem 3.1 *The above algorithm*

(i) runs in $O(n \log n)$ steps,

(ii) produces at most $O(1)$ conformations, each of which is known to be $(\frac{1}{2}+\frac{1}{2}\varepsilon)$-consistent with M with probability $1 - O(1/n)$, and

(iii) produces P as one of the conformations with probability arbitrarily close to 1.

Since the algorithm did not make specific use of P (which is unknown), this means that the algorithm will find any consistent conformation with probability close to 1. Since the algorithm only outputs $O(1)$ conformations, this also means that there are at most $O(1)$ conformations that are consistent with M. Hence, by running the algorithm for a constant number of times, *all* consistent conformations (and only $(\frac{1}{2} + \frac{1}{2}\varepsilon)$-consistent conformations) will eventually be output with probability arbitrarily close to 1.

3.2 Bounds on Total Number of Errors

For the most part, we have focused on the number of errors that can be tolerated per row. In what follows, we briefly mention two results that give tight bounds on the *total* number of errors that can be tolerated.

Theorem 3.2 *If P and Q are sets of n points in $\mathbf{R}^3$ such that $P \not\sim Q$, then $D(P)$ and $D(Q)$ differ in at least $n-4$ positions. This bound is tight in the sense that there exist non-congruent P and Q which differ in precisely $n-4$ distances.*

Proof. The second statement is easier. Let R be a set of $n+1$ labeled points with no non-trivial symmetries. (That is, it is not congruent to any re-labeling of itself.) Furthermore, suppose R has five points r_0, r_1, r_2, r_3, r_4 so that r_0 is equal to the reflection of r_1 through the plane spanned by r_2, r_3, r_4. Set $P = R \setminus \{r_0\}$ and $Q = R \setminus \{r_1\}$. It is easily checked that one may choose such an R so that $P \not\sim Q$ but P and Q differ in only $n-4$ distances.

To prove the first statement, we take two sets P and Q, each consisting of n labeled points and differing in fewer than $n-4$ distances, and we show that $P \sim Q$. Let G denote the graph on $\{1, \ldots, n\}$ in which i and j are joined by an edge if $d(p_i, p_j) \neq d(q_i, q_j)$. Since

G has less than $n-4$ edges, it clearly contains four vertices which form an independent set; it is therefore not difficult to choose an ordering $\{v_1, \ldots, v_n\}$ of the vertices of G so that

(i) $\{v_1, \ldots, v_4\}$ form an independent set, and

(ii) for $i \geq 5$, v_{i+1} has at least as many edges into $\{v_1, \ldots, v_i\}$ as v_i does into $\{v_1, \ldots, v_{i-1}\}$.

Now we claim that

(iii) each v_{i+1} $(4 \leq i \leq n-1)$ has at most $i-4$ edges into $\{v_1, \ldots, v_i\}$.

For if not, then by property (ii) of the ordering, v_j would have at least $i-3$ edges into $\{v_1, \ldots, v_{j-1}\}$ for every $j \geq i+1$; thus G would have at least $(n-i)(i-3) \geq n-4$ edges — a contradiction.

But given (i) and (iii), we can use Lemma 2.1 iteratively to conclude that $P \sim Q$. ∎

Thus, if we are given a purported distance matrix M, there can be at most one conformation whose distance matrix differs from M in at most $(\frac{1}{2}-\varepsilon)n$ entries. The following theorem shows that we can find this conformation with probability arbitrarily close to 1 in linear time. The proof is analogous to, but simpler than, that of Theorem 3.1.

Theorem 3.3 *Fix any $\varepsilon > 0$, and distance matrix M. There is a randomized algorithm that runs in expected linear time and with probability close to 1 returns one of the following results.*

(i) An n-point subset P which differs from M in at most $(\frac{1}{2}-\varepsilon)n$ entries, or

(ii) The (correct) assertion that no such P exists.

4 Sparse Matrices with Errors

Now consider the case in which M is a partial matrix, so that only βn entries per row are given, for some arbitrary constant $\beta > 0$. Let G_M denote the graph whose vertex set is the rows (equivalently, columns) of M, and with i and j joined by an edge if $M_{ij} = M_{ji}$ is given.

To obtain a setting in which it is meaningful to try recovering a finite number of candidate conformations, we need to make several assumptions. First of all, if we only assume that G_M has minimum degree βn, then we cannot even guarantee that it is connected. Also, if we want to allow sparsity below $\frac{1}{2}$ (as we clearly do), then no amount of connectivity will guarantee a finite number of conformations against the corruption model we have been using until now. Consider the case in which G_M consists of two copies of $K_{n/2}$ joined by a matching. This graph has sparsity greater than $\frac{1}{2}$ and vertex-connectivity $\Omega(n)$, but an adversary that corrupts a *single* entry per row can delete the matching, resulting in an infinite number of possible conformations.

However, we can handle arbitrary sparsity under the assumption that G_M is any κ-vertex-connected graph — for a constant κ depending on the sparsity — and that corruptions occur *randomly* with probability $(\frac{1}{2}-\varepsilon)$ at each non-blank entry of M. Thus, in this random model, we can state the performance guarantee of our algorithm as follows.

Theorem 4.1 *For all $\beta, \varepsilon, \delta > 0$, there is a constant κ such that the following holds. Let P be an n-point subset of $\mathbf{R}^3$. Let $D(P)'$ be an incomplete version of its distance matrix $D(P)$ so that at least βn entries per row are present and $G_{D(P)'}$ is κ-vertex-connected. (We* *must also make a non-degeneracy assumption about P that will be exposed in the proof.) If M denotes a partial matrix obtained by arbitrarily corrupting each entry of $D(P)'$ with probability $(\frac{1}{2}-\varepsilon)$, then with probability at least $1-\delta$ we can produce in $O(n \log n)$ time a set of $O(1)$ point sets which includes a rigid copy of P.*

Proof. Set $G = G_M$, and let H be the subgraph of G consisting of those edges on which $D(P)'$ was not corrupted. Note that our algorithm knows the identity of G, but not of H. By G-degree and H-degree, we mean the degree of a vertex in these two graphs. We use d_v to denote the G-degree of a vertex v.

As in the algorithm of the previous section, we will maintain a constant number of partially constructed point sets, of which one will agree with P with high probability. If at any step, our collection of partial point sets no longer contains one that agrees with P, we will say the algorithm has *failed.* We wish to show that the overall probability the algorithm fails is at most $1-\delta$.

As before, we pick a random ordering of the points $V = \{v_1, v_2, \ldots, v_n\}$; let d_i denote the number of G-edges from v_i into $\{v_1, \ldots, v_{i-1}\}$. We choose a constant c large enough so that with probability close to 1,

(i) the subgraph induced on $\{v_1, v_2, \ldots, v_c\}$ has minimum H-degree at least $(\frac{1}{2}+\frac{1}{2}\varepsilon)\beta c$, and

(ii) for all $i \geq c$, $d_{i+1} \geq (1-\frac{1}{2}\varepsilon)\beta i$.

The value of c clearly depends on β^{-1}. Such an ordering will be called a *good* ordering; we may assume that the ordering we have chosen is good and consider how to reconstruct P.

First, let $G_c = G[\{v_1, \ldots, v_c\}]$. We initially produce all subgraphs F of G_c that have minimum degree at least $(\frac{1}{2}+\frac{1}{2}\varepsilon)\beta c$ — since we have a good ordering, one of these will in fact be equal to $H_c = H[\{v_1, \ldots, v_c\}]$. For each of these subgraphs F, we apply Theorem 2.5 to identify at most $r = O(\beta^{-1})$ rigid components, and choose a conformation for each of these components. A key point is that r depends only on β, not on c. We can now state our requirement for the connectivity constant κ: it must be large enough in terms of r and δ so that

$$[e^{(1/4)}(3/2)^{(-3/4)}]^{(\kappa/r)} < \delta(4r)^{-1}.$$

Having done this for H_c, we now enumerate the (constant number of) rigid conformations of these components. As part of this enumeration, we obtain rigid subsets $S_1, \ldots, S_r$, for $r = O(\beta^{-1})$, so that our conformation for each S_i is the same as its conformation in P.

Now we use these r rigid subsets as the seeds around which to grow the rest of our point set, following the good ordering. At stage i, we place vertex v_i rigidly into one of the sets $S_1, \ldots, S_r$. Consider vertex v_i, $i \geq c$. We know that d_i, its G-degree into $\{v_1, \ldots v_{i-1}\}$, is at least $(1-\frac{1}{2}\varepsilon)\beta i$; thus there is some j between 1 and r so that its G-degree into S_j, denoted d'_i, is at least $(1-\frac{1}{2}\varepsilon)\beta i/r$. We can find this j by simple enumeration when $i = O(\log n)$, and otherwise by sampling $O(\log n)$ of the edges from v_i into $\{v_1, \ldots v_{i-1}\}$ and taking the majority — this will succeed with high probability over all i.

Now, which of the d'_i many G-edges from v_i into S_j belong to H is determined by independent random choices of bias $(\frac{1}{2}+\varepsilon)$; thus, there is some constant $\lambda_2 < 1$ so that the probability v_i has fewer than $(\frac{1}{2}+\frac{1}{2}\varepsilon)d'_i$ H-edges into S_j is at most λ_2^i. We will say that our ordering is *H-respecting* if no v_i, $i \geq c$ ever has fewer than $(\frac{1}{2}+\frac{1}{2}\varepsilon)d'_i$ H-edges into its chosen S_j. Thus the probability that our ordering is not H-respecting, given that it is good, it at most $\sum_{i\geq c}\lambda_2^i$, which can be made arbitrarily close to 1 by choosing a sufficiently large value for the constant c.

Thus, with probability arbitrarily close to 1, our ordering is both good and H-respecting. For such an ordering, we can use the technique of the previous section to place each point in V correctly *with respect to its chosen* S_j, with high probability and in time $O(\log n)$ per point. Since all steps of the algorithm up till now can be made to succeed with probability arbitrarily close to 1, we may assume that the probability that the algorithm has already failed is at most $\frac{1}{2}\delta$. Unlike the previous section, however, we have now produced, with probability at least $1-\frac{1}{2}\delta$, a partition of V into r rigid sets $S_1,\ldots,S_r$, each of which agrees with the corresponding subset of P. Thus, in order to obtain P, we have to "stitch together" this constant-size collection of rigid sets.

This is where we make use of our connectivity and non-degeneracy assumptions. The crux of this is, for some pair of sets S_i, S_j, to be able to produce a constant number of conformations for $S_i \cup S_j$ so that one agrees with P with probability at least $1-\delta(2r)^{-1}$. We can then iterate on the resulting set of $r-1$ rigid sets, eventually producing a collection of point sets so that one is congruent to P with probability at least $1-\delta$.

We locate and merge a pair S_i, S_j as follows. We first choose any S_i. Note that for a large enough choice of c, we can be sure by property (i) of the good ordering, and Theorem 2.5, that $|S_i| > \kappa$. Let $\alpha(S_i)$ denote the set of edges with exactly one end in S_i. Since G_M is κ-vertex-connected, there exist κ edges in $\alpha(S_i)$ with mutually distinct endpoints; moreover, a $1/r$ fraction of these all enter the same set S_j, for some $j \neq i$. Let E_{ij} denote this set of S_i-S_j edges, and $e = |E_{ij}|$

Let us use E'_{ij} to denote the set of edges in E_{ij} that belong to H. We claim that, under a non-degeneracy assumption, adding any six of these edges in E'_{ij} to $S_i \cup S_j$ will result in $O(1)$ rigid conformations for $S_i \cup S_j$. This follows from

Claim 4.2 *Let N_{12} denote the graph which is obtained by taking two copies of K_6 and joining them by a matching. (That is, N_{12} has vertex set $\{u_1,\ldots,u_6,v_1,\ldots,v_6\}$, with edges between all (u_i,u_j) pairs, edges between all (v_i,v_j) pairs, and edges (u_i,v_i) for $i=1,\ldots,6$.) Then N_{12} is generically rigid in $\mathbf{R}^3$.*

Proof. The following is a method due to Hennenberg [8] for providing a combinatorial proof for the generic rigidity of a given graph G. Starting with G, we repeatedly apply one of the following two rules:

(i) Delete a vertex of degree 3.

(ii) Delete a vertex of degree 4 and add an edge between two vertices in its former neighbor set.

If this process produces a K_3, then G is generically rigid. Such a sequence of operations (and intermediate graphs) is called a *Hennenberg 3-sequence*.

We in fact prove the stronger result that the graph N'_{12} is generically rigid, where N'_{12} is the graph obtained by deleting the six edges (u_i,u_j) and (v_i,v_j) for $4 \leq i \leq j \leq 6$. We prove this by simply constructing the following Hennenberg sequence:

(1) We eliminate $\{u_4, u_5, u_6\}$ using rule (ii), adding a matching between $\{u_1, u_2, u_3\}$ and $\{v_4, v_5, v_6\}$.

(2) We eliminate $\{v_4, v_5, v_6\}$ using rule (ii), adding a matching between $\{u_1, u_2, u_3\}$ and $\{v_1, v_2, v_3\}$.

(3) The graph we obtain is a copy of K_6 minus a matching, for which it is well-known how to find a Hennenberg sequence. ■

In view of Claim 4.2, we can state our non-degeneracy assumption about P in context — P does not contain 12 points whose interpoint distances on some copy of N_{12} induce a flexible structure. By the definition of generic rigidity, the collection of ordered point sets on which this happens is the complement of a dense open subset of $\mathbf{R}^{12\times 3}$, and thus is a degeneracy.

We make use of this claim as follows. Of the edges between S_i and S_j, if we were to select 6 that belonged to the good set E'_{ij}, then by Claim 4.2, we would obtain $O(1)$ configurations consistent with these 6 measurements, of which one agreed with P. Now, using our definition of κ in terms of r and δ, the probability that fewer than $e/4$ edges of E_{ij} belong to H is at most $\delta(4r)^{-1}$. And if $\left|E'_{ij}\right| \geq e/4$, then we have a constant probability that six given edges all lie in E'_{ij}; by trying a sufficiently large constant number of 6-tuples, we will in the process enumerate a conformation for $S_i \cup S_j$ that agrees with P with probability at least $1 - \delta(4r)^{-1}$. Thus, the overall probability that we fail to merge S_i and S_j successfully is at most $1 - \delta(2r)^{-1}$.

The theorem then follows by iterating on this "merging" construction $r - 1$ times, as argued above. ■

Acknowledgements

We thank Jamie Williamson for valuable discussions on the issues that arise in interpreting NMR data, and Robert Connelly for answering some of our questions about rigidity theory. We would also like to thank Dan Kleitman and Sergey Fomin for helpful discussions.

References

[1] E.D. Bolker, B. Roth, "When is a bipartite graph a rigid framework?", *Pacific J. Math*, 90(1980), pp. 27–44.

[2] B. Bollobas, *Extremal Graph Theory*, Academic Press, 1978.

[3] A.L. Cauchy, "Sur les polygones et polyédres," *J. École Polytechnique*, 19(1813), pp. 87–90.

[4] R. Connelly, "On generic global rigidity," in *Applied Geometry and Discrete Mathematics*, P. Gritzmann and B.Sturmfels, eds., pp. 147–155.

[5] R. Connelly, personal communication, 1996.

[6] G.M. Crippen, T.F. Havel, *Distance Geometry and Molecular Conformation*, Wiley, 1988.

[7] A. Dress, T.F. Havel, "Shortest-path problems and molecular conformation," *Disc. App. Math.*, 19(1988), pp. 129–144.

[8] L. Hennenberg, *Die graphische Statik der starren Systeme*, Leipzig, 1911.

[9] B. Hendrickson, "Conditions for unique graph realizations," *SIAM J. Computing*, 21(1992), pp. 65–84.

[10] B. Hendrickson, *The Molecule Problem: Determining Conformation from Pairwise Distances*, Ph.D. thesis, Cornell University Dept. of Computer Science, 1990.

[11] G. Kolata, "Geodesy: dealing with an enormous computer task," *Science*, 200(1978), pp. 421–422.

[12] A. Lane, T. Fulcher, "^{1}H NMR relaxation and NOE's in nucleic acids," *J. Magnetic Resonance*, 107(1995), pp. 34–42.

[13] J. Milnor, "On the Betti numbers of real varieties," *Proc. American Mathemetical Society* 15(1964), pp. 275–280.

[14] F. Ni, "Two–dimensional transferred NOE effects with incomplete averaging of free- and bound-ligand resonances," *J. Magnetic Resonance*, 106(1995), pp. 147–155.

[15] J. Saxe, "Embeddability of weighted graphs in k-space is strongly NP-hard," *Proc. 19th Allerton Conf. on Computers, Controls, and Communications*, 1979, pp. 480–489.

[16] R. Thom, "Sur l'homologie des variétés algébriques réelles," *Differential and Combinatorial Toplogy*, S.S. Cairns, ed., Princeton University Press, 1965.

[17] A. Wang, S. Kim, P. Flynn, S. Chou, J. Orban, B. Reid, "Errors in RNA NOESY distance measurements in chimeric and hybrid duplexes: differences in RNA and DNA proton relaxation," *Biochemistry*, 31(1992), pp. 3940–3946.

[18] W. Whiteley, "Motions of a bipartite framework," *Pacific J. Math*, 110(1984), pp. 233–255.

[19] G. Wider, S. Macura, A. Kumar, R. Ernst, K. Wüthrich, "Homonuclear two-dimensional ^{1}H NMR of proteins. Experimental procedures," *J. Magnetic Resonance*, 56(1984), pp. 207–234.

[20] J. Williamson, personal communication, 1995.

[21] W. Wunderlich, "Untersuchungen zu einem trilaterations mit komplanaren standpunkten," *Sitz. Osten. Akad. Wiss.*, 186(1977), pp. 263–280.

[22] K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, 1986.

[23] L. Zhu, B, Reid, "An improved NOESY simulation program for partially relaxed spectra: BIRDER," *J. Magnetic Resonance*, 106(1995), pp. 227–235.