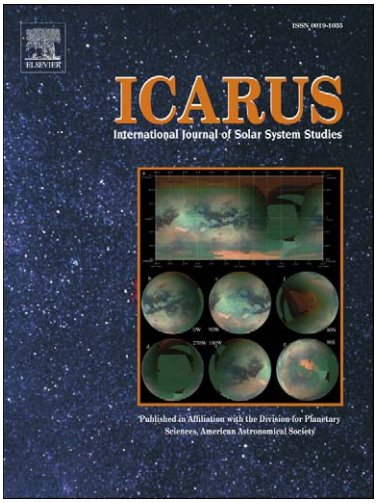
# Asteroid identification over apparitions

Mikael Granvik, Karri Muinonen




# Asteroid identification over apparitions

Mikael Granvik, Karri Muinonen

PII: S0019-1035(08)00230-3
DOI: 10.1016/j.icarus.2008.06.005
Reference: YICAR 8708

To appear in: *Icarus*

Received date: 4 September 2007
Revised date: 18 May 2008
Accepted date: 11 June 2008

Please cite this article as: M. Granvik, K. Muinonen, Asteroid identification over apparitions, *Icarus* (2008), doi: 10.1016/j.icarus.2008.06.005

# Asteroid identification over apparitions

[1,2]Mikael Granvik and [1]Karri Muinonen

E-mail: mgranvik@iki.fi

[1]Observatory, P.O. Box 14, FI-00014 University of Helsinki, Finland
[2]Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822

Submitted to *Icarus*
Submitted 4 September 2007
Revised 28 January 2008
Revised 18 May 2008

Manuscript pages: 28
Tables: 6
Figures: 4

**Proposed Running Head:** Asteroid identification over apparitions

**Editorial correspondence to:**
Mikael Granvik
Insitute for Astronomy
University of Hawaii
2680 Woodlawn Drive
Honolulu, HI 96822
Phone: +1 808 956 0982
Fax: +1 808 988 3893
E-mail: mgranvik@iki.fi

**Abstract**

We present a new method for the linking of scarce asteroid astrometry over apparitions, and apply it both to simulated and real data to prove its feasibility. Up to date, there has not been a robust method available to search for linkages between the approximately 50,000 provisionally designated sets of asteroid astrometry spanning less than two days. Unless such a scarce set of astrometry is linked to another set of astrometry, the underlying object can be considered lost as the ephemeris uncertainties are substantial. The new method, which can tackle the challenges, is based on Ranging, which is a fully nonlinear, statistical orbital inversion method. Ranging properly treats astrometric uncertainties and propagates the uncertainty to the resulting orbital-element probability density, which is sampled by a set of orbits. The new orbital-element-space multiple-address-comparison (oMAC) method uses dimensionality-reduction techniques and tree structures to efficiently search for overlapping probability densities in the orbital-element phase space. Overlapping probability densities indicate a candidate linkage between astrometric observation sets. To accept a candidate linkage, we have to find a many-body orbital solution which reproduces the observed positions within the observational uncertainties. To find the linking orbit, we use a multi-step approach starting from a Monte-Carlo generation of possible orbits in a reduced volume of the orbital-element phase space and ending with a least-squares orbital solution, which, in addition to the Sun's gravitation, also takes into account the gravitational influence of the relevant planets. The new multiple-address-comparison method has a loglinear computational complexity, that is, it scales as $\mathcal{O}(n \log n)$, where $n$ is the number of included observation sets. It has recently also been implemented for the ephemeris-space multiple-address-comparison (eMAC) method, which is optimized for the short-term linking of scarce astrometry.

**Key Words:** ASTEROIDS, DYNAMICS, ORBITS, CELESTIAL MECHANICS

# 1 Introduction

Currently, some 70,000 provisionally designated observation sets in the Minor Planet Center's (MPC) astrometric observation database belong to what we call single-apparition sets (SAS). Roughly 50,000 of the 70,000 SASs span less than 48 hours (hereafter referred to as 48-hour SASs) and most of these span at least two nights, as the official MPC guidelines require to obtain a provisional designation. In practice, most of the underlying objects can be considered lost due to the large ephemeris uncertainties stemming from such scarce data sets. Only if the observation set can be linked to additional astrometry—new or archived—can the resulting uncertainties be reduced to a level where the object is no longer considered lost. The total number of SASs grew for several years, but recently their number has been fluctuating around the above-mentioned 70,000. As the number of discovered objects continues to grow, the percentage of SASs with respect to all discovered objects is currently decreasing. However, the SASs still amount to approximately 15% of all observation sets that have received a provisional designation from the MPC.

It has been assumed that a number of new linkages should be found among the 48-hour SASs themselves (personal communication with T. Spahr). However, up to date there has not been a robust method available to search for linkages between the sets. Specifically, a suitable method should be designed to deal with the extremely short observational time spans and long linking intervals typical for the 48-hour SAS data. Earlier methods—such as the ephemeris-space multiple-address-comparison (eMAC) method by Granvik & Muinonen (2005, hereafter GM05) and the multiple-solution orbit-identification method by Milani et al. (2005b)—might solve parts of the problem but they have some method-specific limitations.

When optimizing, analyzing, and comparing the performance of binary classifiers—of which a linking method is an example—we need two comparison variables. Usually an improvement in the first leads to a degradation in the second, and vice versa. As comparison metrics, we use the sensitivity and the positive-predictive value as described by, e.g., Granvik et al. (2007). The sensitivity $Sens$ is a measure for how well a binary classifier correctly identifies a condition and it is defined as

$$Sens(C) = \frac{l_c}{L_c}\,, \tag{1}$$

where $C$ is the binary classifier, or linking method, $l_c$ is the number of correct linkages detected, and $L_c$ is the number of correct linkages present in the data. The positive-predictive value $PPV$ reflects the probability that a detected

linkage is a correct one and it is defined as

$$PPV(C) = \frac{l_c}{l_c + l_e}\,, \tag{2}$$

where $l_e$ is the number of erroneous linkages detected.

Whereas the eMAC method is, in principle, a general solution to any currently imaginable asteroid linking problem (see, for example, Granvik et al. 2007), it is optimized to deal with short linking intervals. In the eMAC method, candidate linkages are sought by comparing ephemerides at a few common epochs. To make sure most, if not all, correct linkages are detected, that is, to optimize the sensitivity, the computed ephemerides must be accurate. Long linking intervals would therefore require the computationally burdensome propagation of the orbital-element probability-density function (p.d.f.) from the inversion epoch around the observational mid-date to the common comparison epochs using a many-body dynamical model. The many-body dynamical model (hereafter referred to as the $n$-body model) takes into account the gravitational influence by relevant planets in addition to the gravitational influence by the Sun. The $n$-body model could possibly be changed to the considerably easier two-body model—which only accounts for gravitational influence by the Sun—with the probable expense that some of the correct linkages would be missed. Another—probably more severe—drawback is due to the nonlinearities induced into the orbital-element p.d.f. during the propagation from the inversion epoch to the comparison epochs. A substantial increase in the number of sample orbits as compared to short-term linking would be required to make sure that the ephemeris-space is properly sampled at the comparison epochs. In principle, the latter problem could also be prevented by increasing the maximum difference in ephemeris space between accepted linkages. Inaccurate ephemerides would therefore not lead to the rejection of correct linkages. However, the simplistic solution would imply an increase in the number of false-positive candidate linkages, that is, decrease the positive-predictive value, which might render the method useless.

The multiple-solution orbit-identification method assumes a partially-Gaussian orbital-element p.d.f., and can be used for observation sets containing three or more observations. As the orbital uncertainties are properly treated in only one dimension, it is not clear how sensitive the method is for correct linkages given typical astrometric uncertainties and the limited amount of data contained in, for example, 48-hour SASs. To the best of our knowledge, there has not been simulation results published for the multiple-solution method that would allow us to estimate its performance when linking over apparitions. However, the multiple-solution method has successfully

scanned real SAS data and found some 1,500 linkages accepted by MPC, which is considerably more than what was found with single-solution methods using the same data (Milani et al. 2005b). A fraction of the linkages found are linkages between 48-hour SASs, but we assume that the 50,000 48-hour SASs still hold a number of linkages to be found with more accurate methods. The work by Milani et al. (2005b) clearly shows that the more one allows the orbital-element p.d.f. to deviate from a Gaussian distribution, the higher the sensitivity will become. As our linking methods are based on fully non-Gaussian orbital-inversion methods, it is perceivable that we can reach a higher sensitivity as compared to methods based on Gaussian approximations.

With the new large-scale surveys such as the University of Hawaii's Panoramic Survey Telescope And Rapid Response System 1 (Pan-STARRS 1; see, for example, Jedicke et al. 2007) and ESA's astrometry mission Gaia (Mignard et al. 2008)—respectively coming online in 2008 and 2012—it is of utmost importance that new identification methods can cope with challenges such as the substantially increasing data rate, relatively large parallaxes, and long linking intervals. One of the most important factors is the method's scalability, that is, the increase in computing power which is required in order to analyze larger data sets. Up to date, linking methods have typically scaled as $\mathcal{O}(n^2)$ where $n$ is the number of included observation sets, that is, they have a quadratic computational complexity. Doubling the size of a data set has led to a computing time four times longer using the same platform. Recently, Kubica et al. (2007) published a new method for the short-term linking of asteroid astrometry immediately after discovery. The new method scales as $\mathcal{O}(n \log n)$ and uses $k$-dimensional trees to efficiently prune all impossible paths between observed positions. The new method has a so-called loglinear computational complexity. Here we present a loglinear solution for the long-term linking problem. Our technique is essentially based on augmented red-black binary trees and it has also recently been implemented for our short-term linking method (GM05).

Our aim is to present the overall structure of a new, statistical linking method which is specifically designed to link scarce astrometric observation sets over several apparitions, and to apply the linking method to both simulated data and a few real examples to prove its feasibility. The paper is organized as follows. Sect. 2 explains the generation of simulated astrometric data for testing purposes, and presents a few key conclusions which have been used when designing the linking method. In Sect. 3, the new long-term linking method is presented. The results are put forward and discussed in Sect. 4 and, finally, our conclusions are given in Sect. 5.

## 2 Simulations

Simulated observations of main-belt objects (MBOs) and near-Earth objects (NEOs) were generated using the ASurv software (see, for example, GM05). ASurv randomly draws orbital elements and absolute magnitudes from given cumulative distribution functions (c.d.f.) which here were the debiased c.d.f.s by Jedicke et al. (2002). Upper limits of $H = 13\,\mathrm{mag}$ and $H = 20\,\mathrm{mag}$ were used for the absolute magnitude for MBOs and NEOs, respectively. Positions and apparent magnitudes for these random objects were then computed for specified observation dates. If the position fell inside the observation window and the apparent magnitude was lower than a given threshold (here, $V_{\mathrm{lim}} = 14\,\mathrm{mag}$ for MBOs and $V_{\mathrm{lim}} = 20\,\mathrm{mag}$ for NEOs), the simulated observation was accepted. The adopted magnitude limits are chosen to provide us with a sufficient amount of SASs to be able to optimize the sensitivity of the linking method, not to provide an entirely realistic survey simulation resembling future large-scale surveys. The goal was to use realistic orbits and to obtain detections at varying solar elongations. Here we allowed the target to be anywhere on the sky with the only limit being a minimum solar elongation of 45°. The dynamical model used in propagations between observation dates took into account perturbations induced by all planets as well as the dwarf planet Pluto. The leading relativistic term due to the Sun was also included (Sitarski 1983). The positions of the perturbers were extracted from the Jet Propulsion Laboratory's DE405 planetary ephemerides (Standish 1998). Finally, uncorrelated Gaussian random noise with zero mean and a standard deviation of $\sigma = 0.5''$ was added to the simulated observations.

We used a cadence of three simulated observations each separated by one hour on two consecutive nights, which was repeated eight times with a time interval of 700 days, or roughly 23 months. The time interval was deliberately chosen to differ from the typical synodic period between the Earth and an MBO (∼14–18 months) to make sure that a given object would be observed at varying opposition-centered longitudes. For a single object, we thus got a maximum of $3 \times 2 \times 8 = 48$ simulated observations, which was split up into eight two-nighters. The maximum time span of the combined data set is roughly 13 years and 5 months. Note that the chosen cadence is not realistic, because common dates are assumed for the detections. The results obtained for simulated data sets are thus not entirely comparable to the results obtained for real data.

By using five different seed values for the random number generator, we generated five different master sets of simulated MBO observations as well as five different sets of simulated NEO observations (Table 1 and Figs. 1–3). The application of preliminary versions of the linking method to the

simulated data led to three key findings which guided the development of the current method.

Insert Figs. 1–3 here.

First, if searching for 2-linkages, a relatively large fraction ($\sim 10\%$) of the proposed linkages are erroneous. In practical terms, this showed up as $n$-body orbits satisfactorily reproducing erroneously linked astrometry. As an example, Table 2 shows the residuals resulting from the nominal $n$-body orbit (Table 3) which erroneously links two simulated SASs. Interestingly, we found negative eigenvalues for the linearized covariance matrix obtained for the least squares solution erroneously linking the two data sets. The conditioning number—the largest eigenvalue divided by the smallest eigenvalue—of the Fisher information matrix for the orbital elements was $\sim 10^9$. The inverse of the information matrix, that is, the covariance matrix, was obtained both using the LU decomposition algorithm and the Cholesky decomposition algorithm (Press et al. 1999). Both methods yielded essentially identical results. Even though the observational time span is almost ten years, the orbital inverse problem is nonlinear due to the extremely uneven distribution of the astrometry. Focusing on 3-linkages, the number of proposed erroneous linkages fell to virtually zero without a substantial impact on the sensitivity. Note that even though we decided to restrict the search to ($n \geq 3$)-linkages only, the techniques used can be changed to scan the data for 2-linkages.

Second, checking every possible triplet is impossible in practice. The number of different, unordered triplets $n_3$ in a master set containing $n$ subsets is given by the binomial constant $\binom{n}{3}$, which is equal to

$$n_3 = \frac{n(n-1)(n-2)}{6}\,. \tag{3}$$

As the simulated MBO master sets contain on average 1,916 two-night observation sets, the average maximum number of trial 3-linkages between the simulated observation sets is thus 1,170,455,660. The average number of correct 3-linkages is 2,696. For simulated NEOs, the respective numbers are 2,098, 1,536,894,096, and 1,532. For MBOs, roughly two triplets in a million possible 3-linkages are thus correct linkages, whereas for NEOs roughly one triplet in a million is a correct linkage. The current real data set contains $\sim$50,000 48-hour SASs, which means that there exists $\sim 2.1 \times 10^{13}$ possible 48-hour-SAS triplets. Rigorously checking such an amount of triplets is—at least—challenging.

Third, for a large fraction of the simulated correct MBO and NEO 3-linkages, a least-squares orbital solution can fairly satisfactorily explain the data even though a simple two-body dynamical model is used. For 13,327, or 98.9%, of the 13,478 correct MBO 3-linkages, we obtained a two-body least-squares solution with the *Observed minus Computed* ($O - C$) residual

rms (hereafter referred to as the rms) less than 100″, whereas of the 7,692 correct NEO 3-linkages, 7,547, or 98.1%, resulted in an rms value better than 100″ (Fig. 4). For 86 MBO 3-linkages and 29 NEO 3-linkages, we were unable to obtain a successful two-body solution. Note that, for 2-linkages, the percentages would naturally be even more impressive.

Insert Fig. 4 here.

# 3 Methods

The current long-term linking method uses two main filters; the first one—the orbital-element-space multiple-address-comparison (oMAC) filter—is used to find a substantially reduced set of trial linkages (as compared to all possible trial linkages) worth to be analyzed in detail. The second filter tries to find a full $n$-body orbital solution which reproduces the observed astrometry of several combined observation sets given realistic observational uncertainties. If a trial linkage passes the second filter it thus implies a linkage between the sets. To ease the computational load, we have divided the second filter into three subfilters. Whereas the last subfilter comprises a complete $n$-body analysis, the two first subfilters require that the sets also have to be fairly satisfactorily fit using a two-body dynamical model. Linkages between observation sets of objects exhibiting strongly non-Keplerian motion, such as close approaches to planets, during the total observational time interval may therefore be erroneously discarded. A large part of the analysis can be made using the simplistic and computationally efficient two-body dynamical model, because the first main filter also assumes that the orbital motion is Keplerian, or at least nearly-Keplerian. For the motivation, we refer the reader to the results obtained in Sect. 2. Note that as the computers become faster (and/or processor cores more numerous), we may choose to remove the two-body approximation in the future and use the full $n$-body dynamical model throughout the analysis.

Before going into details of the linking algorithm, we will shortly describe the orbital-inversion methods used.

## 3.1 Orbital inversion in constrained phase-space volumes

For scarce observation sets, a rigorous sampling of the orbital-element p.d.f. is critical to be able to link observation sets over long time intervals. For the inversion of the SASs for the orbital-element p.d.f.s, we use either statistical orbital ranging (Ranging; Virtanen et al. 2001, Muinonen et al. 2001), or the phase-space Volumes-of-Variation method (VoV; Muinonen et al. 2006). Typ-

ically, Ranging is used in the domain before the so-called phase transition in the orbital uncertainty (Virtanen et al. 2005, Muinonen et al. 2006), whereas VoV is optimized for the phase-transition domain. Least squares with linearized covariances (LSL; see, for example, Muinonen & Bowell 1993) is used after the phase transition when the inverse problem can typically be treated with linearized methods. The observational time intervals to reach the LSL domain are usually weeks for NEOs and months for MBOs. Typically, we therefore use Ranging or VoV on 48-hour SASs, and LSL on observation sets combined of two or more SASs. Note that even if the orbital-element covariance matrix can fail to correctly represent the orbital uncertainties for combinations of two SASs (see Sect. 2), the nominal least-squares orbit can still be useful. The inversion epoch of each observation set is here defined as the midnight (TT) closest to the mid-date of the observations.

The orbital-element space to be explored by Ranging or VoV sampling can be constrained by using an informative a priori p.d.f. In practice, one can, for example, focus only on NEOs and the so-called inner-Earth objects (IEOs)—that is, objects whose orbits lie completely within the Earth's orbit—and thus require that the perihelion distance $q = a(1 - e)$ of acceptable sample orbits must be less than 1.3 AU.

In addition to the standard version of LSL, we also make use of the well-known incomplete differential correction technique. During incomplete differential correction, one or more elements are fixed when correcting the rest. To tackle severe nonlinearities, we also have the option to make partial correction steps. When using partial steps one does not correct the orbital elements by the correction computed, but use only, say, a tenth of the correction in each element. Strong nonlinearities will therefore—hopefully—not lead the solution astray.

## 3.2 Orbital-element-space multiple-address comparison

The oMAC filter is similar to the ephemeris-space multiple-address-comparison (eMAC) filter presented in GM05 with the difference that orbital elements are used as comparison variables and the comparison algorithm has been further developed to scale as $\mathcal{O}(n \log n)$ where $n$ is the number of observation sets included. Essentially, the first filter requires that the p.d.f.s of the orbital elements computed from the three separate observation sets have to overlap, in general at a specified comparison epoch.

We tried several different comparison variables such as Cartesian orbital elements, Keplerian orbital elements, equinoctial orbital elements, Poincaré variables, the angular momentum vector, heliocentric spherical coordinates, and a few of their combinations. The number of comparison epochs was

also altered when spherical coordinates were tested. When using five of the Keplerian orbital elements—the semimajor axis $a$, eccentricity $e$, inclination $i$, longitude of the ascending node $\Omega$, and the argument of perihelion $\omega$—and the time of ascending node $t_\Omega$ folded by the orbital period to be as close to 1.0 January 2000 as possible, we found that oMAC's positive-predictive value for MBOs was about three times higher than when using the other types of comparison variables. For NEOs, the improvement in oMAC's positive-predictive value using the same comparison variables was only about 10%–20%. The time of ascending node was chosen instead of the mean anomaly $M_0$ (or the time of perihelion $\tau$) because the plane of the orbit is typically fairly well determined. Note that a two-body dynamical model is assumed, and we do not integrate the orbital elements but use fast analytical methods for their propagation in time.

In practice, we compute a sufficient amount, say 10,000, discrete orbits that sample the true orbital-element p.d.f. using Ranging or VoV. The probabilistic treatment is not exact, because we assume an equal weight for each sample orbit. The proper weight would be based on the $\chi^2$ of the residuals, relevant Jacobian terms, and a term securing the invariance in parameter transformations. The simplified treatment—based solely on residual intervals—does not greatly affect the extent of the orbital-element p.d.f. in the phase space. As we are not currently using the relative weights of the sample orbits in the oMAC method, their computation would result in unnecessary complications and consumption of computational resources. However, we do not want to rule out the utilization of the weights in future developments.

For the comparison phase, the six-dimensional comparison vector ($a$, $e$, $i$, $\Omega$, $\omega$, $t_\Omega$) is squeezed into a single integer, that is, an address, which permits fast comparison of different orbits (for details, see GM05 and Muinonen et al. 2005). Discretizing each dimension of a hypervolume of a $D$-dimensional space into $m_d$ ($d = 1, 2, \ldots, D$) intervals, the total number of allowed addresses (bins) becomes

$$A = m_1 m_2 \cdots m_D. \tag{4}$$

Let the $D$ indices of a certain bin be $i_d > 0$ ($d = 1, 2, \ldots, D$). That given bin then obtains an address given by the single integer

$$I = 1 + \sum_{j=1}^{D} (i_j - 1) \prod_{k=1}^{j-1} m_k. \tag{5}$$

Of the tens of different discretizations tested, the one described in Table 4 provided the best combination of sensitivity and positive-predictive value.

However, we cannot rule out that even better discrectizations could be found in the future.

Once identical addresses are found for sample orbits originating from three separate observation sets, we conclude that the observation sets give rise to similar orbits thus implying a potential 3-linkage. The maximum deviation of the three orbits in comparison space is explicitly given by the bin sizes. Whereas GM05 used a quadratic comparison algorithm to find identical addresses, we have developed a new loglinear algorithm which does not directly search for identical addresses, but organizes the addresses so that identical ones can be trivially extracted. The new algorithm is based on red-black (RB) binary trees and circular doubly-linked lists (for descriptions of these data structures, see, for example, Cormen et al. 2003). Given that the orbital-element p.d.f.s resulting from $n$ observation sets are sampled with $m$ orbits, we start by compressing the $nm$ six-dimensional orbital-element sets to $nm$ scalars, or addresses, which scales as $\mathcal{O}(nm)$. These addresses are then inserted to an RB tree so that the addresses are used as keys. The identifier of the observation set—that is, the provisional designation or the number—from which an address has been computed is inserted into a circular doubly-linked list within the tree node. If an identifier has already been inserted to a list, it is not inserted again. The insertion to a node scales as $\mathcal{O}(1)$. As the insertion to RB trees scales as $\mathcal{O}(\log k)$, where $k$ is the number of nodes, the insertion of $nm$ addresses is thus guaranteed to scale as $\mathcal{O}(nm \log nm)$. After the insertion process, each node of the RB tree contains a list of identifiers for observation sets that lead to an identical address in the orbital-element space. In other words, all observation sets indicated by a list can possibly be linked and should hence be accepted by the first filtering.

The next step is to try to find a single orbital solution tying together both observation sets corresponding to a potential linkage assuming realistic observational uncertainties.

### 3.3 Linking orbits through Monte Carlo sampling and least squares

As indicated in Sect. 3.1, the preferable method when verifying candidate linkages, that is, when trying to find the linking orbit between SASs in the second filter is the least-squares (LSL) method. Whereas the inversion of scarce SASs of observations typically produces wide orbital-element p.d.f.s (Ranging suitable), the inversion of a combination of two or more SASs separated by several years results in very constrained p.d.f.s so that LSL becomes suitable for finding a single orbit which reproduces the astrometry observed.

As stated in Sect. 2, the orbital uncertainty estimates resulting from a combination of two SASs only are not necessarily reliable if derived using LSL. Typically, the difference in the orbital uncertainties resulting from a single SAS and a set combined of three SASs are several orders of magnitude. It is thus clear that using a sample orbit computed for one of the SASs as an initial orbit for LSL may fail to converge, simply because the initial orbital elements are too far from the final ones and the linearity assumption does not apply.

The sample orbits computed for the separate SASs provide a good first approximation and should hence be used. We use the orbital information of the SASs gathered by the oMAC filter in the form of addresses in the phase-space of the orbital elements. First, we select an address and require that at least three different SASs have obtained that address. Similarly, if 2-linkages were sought, we would require that at least two SASs would have obtained the same address. Note that the choice to search for 3-linkages only stems from the need to increase the method's positive-predictive value. In other words, there are no technical details present which would make it impossible to search for, for instance, 2-linkages or 4-linkages. The original indices $i_d$ can now be retrieved from the address $I$ with the following recurrence relation:

$$i_d = (1 - \delta_{1d}) + \mathrm{int}\left[\frac{I - \sum_{j=d+1}^{D}(i_j - 1)\prod_{k=1}^{j-1} m_k}{\prod_{j=1}^{d-1} m_j}\right], \tag{6}$$

where $\delta_{ij}$ is the Kronecker symbol (note that there is a typo in Eq. (14) in GM05). Using the indices $i_d$ and the bin sizes, we compute the intervals for $a$, $e$, $i$, $\Omega$, $\omega$, and $t_\Omega$.

From the fairly compact region defined by the orbital-element intervals, we draw orbital-element sets for a given epoch in a Monte Carlo (MC) fashion. Hence we call it the two-body MC (2bMC) subfilter. We have chosen the observational mid-date of all SASs having the address under scrutiny as the epoch for the generated orbital elements. For each randomly drawn orbit, we compute the $O - C$ residuals (hereafter referred to as the residuals) with respect to all the SASs having the address under scrutiny. If three or more SASs have all residuals $\epsilon$ smaller than preset residual limit $\epsilon_{\mathrm{2bMC}}$, each of the resulting triplets will pass the 2bMC subfilter. The number of triplets is given by Eq. 2 where $n$ is the number of SASs connected with the MC orbit.

The number of addresses is typically large. The addressing gives us a possibility to prioritize the analysis of specific volumes of the orbital-element phase space. For example, if there are hypothesized populations in the phase space (for example, for some dynamical reasons), we could analyze the corresponding addresses first instead of analyzing the phase space randomly, or

we can optimize the order in which addresses are analyzed by using an a priori distribution. The a priori distribution can be constructed using, for example, the ASTORB database which contains osculating elements for all numbered or designated asteroids (Bowell et al. 1994). Using the discretization parameters in Table 4, an address can be computed for each of the orbits in the database. The intersection of the original addresses and the addresses of the known population as derived from the ASTORB database then leaves us with addresses corresponding to orbits that have already been detected. Finally, the resulting addresses can be sorted in descending order based on their frequency in the known population. Acceptable linkages are now most probably found in the most frequent addresses, and the most likely linkages are thus found first. Note that the use of the a priori distribution still requires that all addresses are analyzed in order to detect all correct linkages. According to preliminary results, using the a priori helps us to detect linkages faster as compared to a random selection of addresses. Note that the rate of linkages found will naturally stagnate at some point.

In the next subfilter—the two-body LSL (2bLSL) subfilter—we try to compute a two-body LSL solution in Keplerian orbital elements for each triplet. A successful two-body solution and rms smaller than a preset limit $\epsilon_{\mathrm{rms,2bLSL}}$ are the requirements to pass the 2bLSL subfilter. We first attempt a complete differential correction and, if unsuccessful, we try to find a solution by consequently applying three different versions of the incomplete differential-correction technique before again applying the complete differential correction. In the first incomplete differential correction we fix all elements but the semimajor axis, in the second we fix all elements but the semimajor axis and the eccentricity, and in the third we only fix the inclination and the longitude of the ascending node. For all cases, that is, both complete and incomplete differential corrections, we use a partial correction step equalling on tenth of the computed nominal correction in each element.

Finally the remaining triplets are scrutinized with a full $n$-body LSL (nbLSL) subfilter. The final acceptance for a triplet is given if the $n$-body LSL residuals conform to the assumed observational uncertainties. In practice, we require that the rms is less than the preset limit $\epsilon_{\mathrm{rms,nbLSL}}$. Again, we first attempt a complete differential correction and if not successful, we resort to the incomplete differential correction before attempting another complete correction. We only use one round of incomplete differential correction were all elements but the semimajor axis and the eccentricity are fixed. For the complete differential corrections we use full correction steps whereas for the incomplete differential correction we use partial correction step which equals on tenth of the computed nominal correction in each element.

Note that at this stage we cannot necessarily be completely certain about

the correctness of the 3-linkages found. A detailed, statistical analysis of the residuals can potentially reveal erroneous linkages, but verifying the correctness of proposed linkages should preferably be done using additional data, new or archive.

Calculating the computational complexity of the second main filter as a function of the number of included observation sets $n$ is nontrivial, because the method is working in address space. The upper limit for the number of addresses is $nm$, where $m$ is the number of sample orbits used for the oMAC filter. A maximum number of addresses would mean that none of the sets can be linked with each other, and the analysis would end (almost) immediately. If only one address is obtained, the computation of residuals in the 2bMC subfilter scales linearly with $n$, and the scaling of the 2bLSL and nbLSL subfilters depends on how many linkages can be found in the data.

# 4 Results and discussion

The new linking method was tested both using simulated data (the generation of which was described in Sect. 2) and by using real data. In the orbital-inversion procedure, the maximum residuals $\epsilon$ for acceptable orbits in both R.A. and Dec. were $|\epsilon| < 6 \times \sigma$, where $\sigma$ is the estimated uncertainty of the astrometry. For the simulated data, we used an uncorrelated astrometric uncertainty of $\sigma_{\mathrm{R.A.}} = \sigma_{\mathrm{Dec.}} = 0.5''$.

By constraining the solution of the orbital inverse problem to relevant phase-space volumes, the oMAC method can be used efficiently. The advantage is that, due to the limited phase-space volume, a substantially reduced number of sample orbits are required to reach the same sensitivity level as compared to an unconstrained solution. A small number of sample orbits in a limited region means a small number of addresses, and the analysis can therefore proceed fast. For an even more detailed search for specific groups or families, we could use a hypothesized a priori distribution during the orbital inversion in order to make sure that the relevant phase-space volume is sampled densely enough. To densify the sampling, we have here constrained the solution of the orbital inverse problem by using two informative a priori p.d.f.s. The first a priori p.d.f. (hereafter the MBO a priori p.d.f.) required that acceptable sample orbits have $a < 5.5\,\mathrm{AU}$ and $q > 1.3\,\mathrm{AU}$, whereas the second a priori p.d.f. (hereafter the NEO a priori p.d.f.) required that acceptable sample orbits have $0.00465424\,\mathrm{AU} < a < 5.5\,\mathrm{AU}$ — the lower limit stems from the radius of the Sun—and $q < 1.3\,\mathrm{AU}$. Note that the MBO a priori allows for Jupiter Trojans and that the NEO a priori p.d.f. allows IEO orbits, sun-grazing orbits, and orbits leading to an impact with the Sun. The

MBO a priori p.d.f. was applied during the inversion of the simulated MBO astrometry, whereas the NEO a priori was applied to the simulated NEO astrometry.

When 10,000 sample orbits had been generated from each simulated SAS, the orbits were fed into the linking method. The values used for the relevant parameters are given in Tables 4 and 5. The average sensitivities for the simulated MBOs and NEOs are approximately 95% and 83%, respectively (Table 6). Noting that the number of NEO orbits included equals approximately one fifth of all known real NEO orbits, the sensitivity estimate should be fairly trustworthy for the real situation as well. The sensitivity for NEOs can be increased to approximately 92% by changing the discretization (discard $t_\Omega$ and reduce the bin size for $\omega$ to 5°), but the cost is an approximately 50% increase in the number of addresses to process after the oMAC filter.

The positive-predictive values computed from the analysis of the simulated data sets are not compatible with the positive-predictive values to be obtained during the analysis of, for example, future large-scale surveys, because the characteristics of the simulated data sets do not accurately resemble the characteristics of real data. For example, because of the fixed cadence used in the simulations, the 2bLSL subfilter can rarely, if ever, be passed by erroneous linkages between different SASs from the same apparition. Furthermore, empirical tests have shown that the number of erroneous linkages found grows quadratically with an increasing sky-plane density of detected objects (GM05 and Milani et al. 2005a). In the analysis of real data and/or a larger amount of more realistic simulated data, the positive-predictive values are therefore expected to be somewhat lower.

The real SASs extracted for testing purposes were the same as the ones given as examples of proposed identifications by Milani et al. (2005b, p. 740):

| | | | | |
|---|---|---|---|---|
| 1992 $SB_3$ | = | 2000 $PG_{14}$ | = | 2003 $GU_{11}$ |
| 1996 $VC_{13}$ | = | 1998 GR | = | 2002 $CB_{303}$ |
| 1999 $DT_4$ | = | 2000 $LC_4$ | = | 2001 $US_{181}$ |
| 1995 $SJ_{32}$ | = | 1998 $MN_{15}$ | = | 2000 $YE_{89}$ |
| 1998 GX | = | 1996 $VJ_{21}$ | = | 2000 $TU_{71}$ |

Except for the last case, the data for the two first SASs in each identification chain have been obtained on two nights and are therefore often called two-nighters. Note that a two-nighter is not necessarily a 48-hour SAS, as the observational time span can be longer for a two-nighter. However, all SASs of the last identification chain are 48-hour SASs. 1998 GX is a two-nighter, 1996 $VJ_{21}$ a three-nighter, and 2000 $TU_{71}$ a single-nighter with an observational time span of only 43 minutes. We used the MBO a priori p.d.f. during the orbital inversion of all SASs, because all the corresponding objects are MBOs.

For the astrometric uncertainty, we assumed $\sigma_{\mathrm{R.A.}} = \sigma_{\mathrm{Dec.}} = 1.0''$. The linking parameters were the same as the ones used throughout the present paper (see Tables 4 and 5) with the exception that for 2000 $\mathrm{TU}_{71}$ we used 50,000 sample orbits in order to keep the sensitivity on a high level despite the short observational time span. The SASs for each of the five 3-linkages were treated simultaneously, and the method successfully detected all five 3-linkages. Note that increasing the number of sample orbits for observation sets with shorter observational time spans can also be used in the opposite direction; we may choose to reduce the number of sample orbits for observation sets having time spans longer than 24 hours and thus reduce the overall computing time without sacrificing the efficiency.

# 5 Conclusions

We have presented a new statistical method for the long-term linking of scarce asteroid astrometry. Up to date, a general method suitable for the task has not been available. The new method is based on nonlinear orbital-inversion methods, which allow us to properly account for observational uncertainties. Due to long linking intervals, the proper treatment of the uncertainties is of utmost importance in order to find the maximum number of correct linkages. The new method has been successfully tested with both simulated and real data.

The present study may have implications for NEO survey strategies as we have shown that linkages over apparitions between scarce sets of astrometry can be detected with a reasonably high sensitivity. In addition to routinely searching for asteroid identifications among scarce data sets, we can think of several additional areas of application for the new method. In our minds, the most important additional application of the method would be the quick scan for additional astrometry when an object potentially impacting with the Earth has been discovered. Using a precomputed set of addresses, the confirmed NEOs as well as NEO candidates with scarce astrometry could be scanned within, say, minutes.

In the future, we will apply the method to observation sets of lost objects such as the 48-hour SASs. The analysis should also include the unidentified single-night sets archived at the MPC.

## Acknowledgments

Research supported, in part, by the Academy of Finland and by the University of Helsinki three-year research grant. MG has also been supported by the foundations of A. Kordelin and M. Ehrnrooth. We thank the two reviewers, A. Milani and T. Spahr, for their constructive criticism.

## References

Bowell, E., Muinonen, K. & Wasserman, L. H. (1994), A Public-Domain Asteroid Orbit Data Base, *in* A. Milani, M. di Martino & A. Cellino, eds, 'Asteroids, Comets, Meteors 1993', IAU Symposium #160, Co-published by IAU and Kluwer Academic Publishers, Dordrecht, p. 477.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2003), *Introduction to Algorithms*, 2 edn, MIT Press, Cambridge, Massachusetts, USA.

Granvik, M. & Muinonen, K. (2005), 'Asteroid identification at discovery', *Icarus* **179**(1), 109–127.

Granvik, M., Muinonen, K., Jones, L., Bhattacharya, B., Delbó, M., Saba, L., Cellino, A., Tedesco, E., Davis, D. & Meadows, V. (2007), 'Linking Large-Parallax Spitzer-CFHT-VLT Astrometry of Asteroids', *Icarus* **192**(2), 475–490.

Jedicke, R., Larsen, J. & Spahr, T. (2002), Observational Selection Effects in Asteroid Surveys and Estimates of Asteroid Population Sizes, *in* W. Bottke, A. Cellino, P. Paolicchi & R. P. Binzel, eds, 'Asteroids III', University of Arizona Press, pp. 71–87.

Jedicke, R., Magnier, E. A., Kaiser, N. & Chambers, K. C. (2007), The next decade of Solar System discovery with Pan-STARRS, *in* A. Milani, G. B. Valsecchi & D. Vokrouhlický, eds, 'Near Earth Objects, our Celestial Neighbors: Opportunity and Risk', IAU Symposium #236, Cambridge University Press, Cambridge, UK.

Kubica, J., Denneau, L., Grav, T., Heasley, J., Jedicke, R., Masiero, J., Milani, A., Moore, A., Tholen, D. & Wainscoat, R. J. (2007), 'Efficient intra- and inter-night linking of asteroid detections using kd-trees', *Icarus* **189**(1), 151–168.

Mignard, F., Cellino, A., Muinonen, K., Tanga, P., Delbó, M., Dell'Oro, A., Granvik, M., Hestroffer, D., Mouret, S., Thuillot, W. & Virtanen, J. (2008), 'The Gaia mission: Expected applications to asteroid science', *Earth, Moon, and Planets* **101**(3–4), 97–125.

Milani, A., Gronchi, G. F., Knežević, Z., Sansaturio, M. E. & Arratia, O. (2005a), 'Orbit determination with very short arcs. II Identifications', *Icarus* **179**(2), 350–374.

Milani, A., Sansaturio, M. E., Tommei, G., Arratia, O. & Chesley, S. R. (2005b), 'Multiple solutions for asteroid orbits: Computational procedure and applications', *Astron. Astrophys.* **431**, 729–746.

Muinonen, K. & Bowell, E. (1993), 'Asteroid orbit determination using Bayesian probabilities', *Icarus* **104**(2), 255–279.

Muinonen, K., Virtanen, J. & Bowell, E. (2001), 'Collision probability for Earth-crossing asteroids using orbital ranging', *Cel. Mech. Dyn. Astron.* **81**(1–2), 93–101.

Muinonen, K., Virtanen, J., Granvik, M. & Laakso, T. (2005), Asteroid orbits with Gaia: inversion and prediction, *in* M. Perryman, ed., 'Three-Dimensional Universe with Gaia', Special Publications SP-576, ESA, Noordwijk, pp. 223–230.

Muinonen, K., Virtanen, J., Granvik, M. & Laakso, T. (2006), 'Asteroid orbits using phase-space volumes of variation', *Mon. Not. R. Astron. Soc.* **368**(2), 809–818.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1999), *Numerical Recipes in Fortran 90 — The Art of* Parallel *Scientific Computing*, 2 edn, Cambridge University Press.

Sitarski, G. (1983), 'Effects of general relativity in the motion of minor planets and comets', *Acta Astronomica* **33**(2), 295–304.

Standish, E. M. (1998), JPL Planetary and Lunar Ephemerides, DE405/LE405, Interoffice Memorandum 312.F-98-048, Jet Propulsion Laboratory.

Virtanen, J., Muinonen, K. & Bowell, E. (2001), 'Statistical Ranging of Asteroid Orbits', *Icarus* **154**(2), 412–431.

Virtanen, J., Muinonen, K., Granvik, M. & Laakso, T. (2005), Collision orbits and phase transition for 2004 $AS_1$ at discovery, *in* Z. Knežević & A. Milani, eds, 'Dynamics of Populations of Planetary Systems', IAU Colloquium #197, pp. 239–248.

Simulated astrometry

| Sim | MBO | | | | NEO | | | |
|---|---|---|---|---|---|---|---|---|
| | SAS | Obj | Obj3 | C3L | SAS | Obj | Obj3 | C3L |
| #1 | 1994 | 696 | 344 | 3013 | 2216 | 1289 | 235 | 1476 |
| #2 | 1878 | 680 | 325 | 2564 | 2099 | 1191 | 227 | 1693 |
| #3 | 1948 | 687 | 350 | 2701 | 2047 | 1213 | 200 | 1531 |
| #4 | 1881 | 680 | 327 | 2498 | 2078 | 1178 | 220 | 1537 |
| #5 | 1879 | 652 | 331 | 2702 | 2049 | 1166 | 228 | 1455 |
| Total | 9580 | 3395 | 1677 | 13478 | 10489 | 6037 | 1110 | 7692 |
| Average | 1916 | 679 | 333 | 2696 | 2098 | 1207 | 222 | 1538 |

Table 1: Parameters describing the simulated MBO and NEO astrometry. *Sim* is the label for a realization of simulated data, *SAS* is the number of SASs in a realization, *Obj* is the number of different objects detected at least once in the realization, *Obj3* is the number of different objects detected at least three times in the realization, and *C3L* is the maximum number of correct 3-linkages to be detected in the data.

$O - C$ residuals for an erroneous linkage

| Set #1 | | | Set #2 | | |
|---|---|---|---|---|---|
| Date | RA [″] | Dec [″] | Date | RA [″] | Dec [″] |
| 2007/11/20 | 0.1 | 0.6 | 2017/06/20 | -0.4 | -0.2 |
| 2007/11/20 | -0.1 | 0.2 | 2017/06/20 | 0.3 | -0.2 |
| 2007/11/20 | 0.1 | -0.8 | 2017/06/20 | 0.4 | 0.2 |
| 2007/11/21 | -0.5 | 0.1 | 2017/06/21 | -0.4 | 0.3 |
| 2007/11/21 | -0.0 | -0.5 | 2017/06/21 | 0.3 | 0.0 |
| 2007/11/21 | 0.4 | 0.4 | 2017/06/21 | -0.2 | -0.0 |

Table 2: Residuals stemming from an $n$-body orbit (see Table 3) erroneously linking two different sets of simulated astrometry. Set #1 has been generated using orbit #1 and set #2 using orbit #2. The Gaussian noise added to the simulated astrometry has a standard deviation of 0.5″. Note that the time interval is roughly 10 years.

Orbits associated with an erroneous linkage

| | $a$ [AU] | $e$ | $i$ [°] | $\Omega$ [°] | $\omega$ [°] | $M_0$ [°] |
|---|---|---|---|---|---|---|
| Orbit #1 | 2.96 | 0.005 | 1.98 | 69.09 | 41.29 | 310.54 |
| Orbit #2 | 3.15 | 0.182 | 1.29 | 114.77 | 52.76 | 33.96 |
| Linking orbit | 3.25 | 0.363 | 3.87 | 52.88 | 9.75 | 162.80 |

Table 3: The Keplerian elements for the two $n$-body orbits which were used when generating the simulated astrometry, and the $n$-body orbit erroneously linking the simulated astrometry. The elements are given for the epoch 2004 Jan 20.25 (TT). Note that even though a good fit in terms of residuals was obtained (Table 2), the linking orbit is substantially different from the true orbits.

Key parameters for the first main filter

| | $a$ | $e$ | $i$ | $\Omega$ | $\omega$ | $t_\Omega$ |
|---|---|---|---|---|---|---|
| Lower limit | 0.0 AU | 0.0 | 0.0° | 0.0° | 0.0° | MJD 50206.0 |
| Upper limit | 5.5 AU | 1.0 | 180.0° | 360.0° | 360.0° | MJD 52906.0 |
| Bin size | 0.1 AU | 0.1 | 0.5° | 10.0° | 45.0° | 180.0 d |
| Bin number | 55 | 10 | 360 | 36 | 8 | 15 |

Table 4: Phase-space discretization parameters for the oMAC algorithm which were empirically found to give the best combination of sensitivity and positive-predictive power. The upper and lower limits for $t_\Omega$ stem from the orbital period of $\lesssim 7$ yr for an object with $a = 5.5$ AU. Note that the resulting number of addresses in the numbered, six-dimensional orbital-element phase space is equal to 3,421,440,000, and that the resolution is highest in $a$ and $i$.

Key parameters for the second main filter

| | |
|---|---|
| $\epsilon_{\mathrm{2bMC}}$ | 4° |
| $n_{\mathrm{orb,NEO}}$ | 100,000 |
| $n_{\mathrm{orb,MBO}}$ | 10,000 |
| $\epsilon_{\mathrm{rms,2bLSL}}$ | 100″ |
| $\epsilon_{\mathrm{rms,nbLSL}}$ | 1.5″ |

Table 5: Key parameters for the second main filter. $\epsilon_{\mathrm{2bMC}}$ is the maximum sky-plane residual allowed in order to pass the 2bMC subfilter, and $n_{\mathrm{orb,NEO}}$ and $n_{\mathrm{orb,MBO}}$ are the maximum numbers of orbits to be generated in the same subfilter. $\epsilon_{\mathrm{rms,2bLSL}}$ and $\epsilon_{\mathrm{rms,nbLSL}}$ are the maximum allowed residual rms values allowed in order to pass the 2bLSL and nbLSL subfilters, respectively.

Linking results for simulated astrometry

| Sim | oMAC | | | | 2bMC | | | | 2bLSL | | | | nbLSL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C3LD | F3LD | Sens [%] | PPV [%] | C3LD | F3LD | Sens [%] | PPV [%] | C3LD | F3LD | Sens [%] | PPV [%] | C3LD | F3LD | Sens [%] | PPV [%] |
| MBO1 | 3,000 | 31,714 | 99.6 | 8.6 | 2,979 | 5,965 | 98.9 | 33.3 | 2,958 | 1,134 | 98.2 | 72.3 | 2,911 | 0 | 96.6 | 100.0 |
| MBO2 | 2,550 | 25,569 | 99.4 | 9.0 | 2,527 | 4,781 | 98.6 | 34.6 | 2,505 | 868 | 97.7 | 74.3 | 2,448 | 0 | 95.5 | 100.0 |
| MBO3 | 2,661 | 31,145 | 98.5 | 7.9 | 2,623 | 5,792 | 97.1 | 31.2 | 2,591 | 1,087 | 95.9 | 70.4 | 2,540 | 0 | 94.0 | 100.0 |
| MBO4 | 2,484 | 27,621 | 99.4 | 8.3 | 2,472 | 5,281 | 99.0 | 31.9 | 2,454 | 1,179 | 98.2 | 67.5 | 2,403 | 0 | 96.2 | 100.0 |
| MBO5 | 2,695 | 27,235 | 99.7 | 9.0 | 2,673 | 5,227 | 98.9 | 33.8 | 2,641 | 924 | 97.7 | 74.1 | 2,591 | 0 | 95.9 | 100.0 |
| NEO1 | 1,423 | 1,379,189 | 96.4 | 0.1 | 1,323 | 28,044 | 89.8 | 4.5 | 1,285 | 92 | 87.2 | 93.3 | 1,250 | 0 | 84.9 | 100.0 |
| NEO2 | 1,611 | 1,095,823 | 95.2 | 0.1 | 1,490 | 22,849 | 88.0 | 6.1 | 1,460 | 66 | 86.2 | 95.7 | 1,440 | 0 | 85.1 | 100.0 |
| NEO3 | 1,420 | 1,107,858 | 92.7 | 0.1 | 1,297 | 20,027 | 84.7 | 6.1 | 1,244 | 63 | 81.3 | 95.2 | 1,219 | 0 | 79.2 | 100.0 |
| NEO4 | 1,472 | 1,076,529 | 95.8 | 0.1 | 1,345 | 17,922 | 87.5 | 7.0 | 1,306 | 49 | 85.0 | 96.4 | 1,283 | 0 | 83.5 | 100.0 |
| NEO5 | 1,382 | 1,017,281 | 95.0 | 0.1 | 1,263 | 19,816 | 86.8 | 6.0 | 1,233 | 77 | 84.7 | 94.1 | 1,213 | 0 | 83.4 | 100.0 |

Table 6: Linking results for the simulated MBO (above) and NEO (below) astrometry. *Sim* is the label for a realization of simulated data, *oMAC* refers to the oMAC filter (see Table 4), *2bMC* to the 2-body MC subfilter (all residuals $< 4°$), *2bLSL* to the 2-body LSL subfilter (residual rms $< 100''$), and *nbLSL* to the full $n$-body LSL subfilter (rms $< 1.5''$). *3LD* is the total number of 3-linkages detected, that is, including both correct and erroneous links, and *C3LD* is the number of correct 3-linkages detected. *Sens* is the sensitivity for correct linkages, and *PPV* is the positive predictive value. Note that the positive-predictive values are strongly case dependent and therefore only portray the results for the current simulations. Note also that the sensitivity is given as the cumulative result of the filtering process with the first filter to the left and the last filter to the right.

Figure 1. Stacked histogram of the arithmetic mean of the solar elongations for simulated SAS included in the analysis. White columns refer to MBOs, whereas black columns refer to NEOs.

Figure 2. Stacked histogram of the rate of motion in ecliptic longitude for simulated SASs of MBOs (white) and NEOs (black). Note that the scale is logarithmic.

Figure 3. As Fig. 2, but for the rate of motion in ecliptic latitude.

Figure 4. The fraction of two-body least-squares solutions having an $O - C$ residual rms value smaller than a given rms value (cf. cumulative density function). Note that the fraction does not reach unity, because the two-body least-squares solution could not be found for 86 MBOs and 29 NEOs.

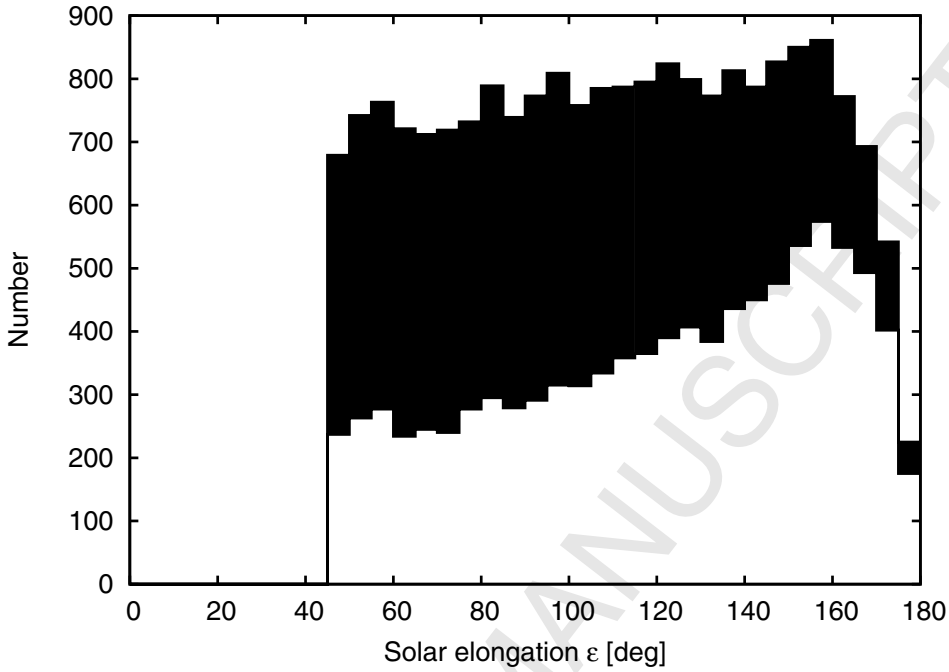



Figure 1:

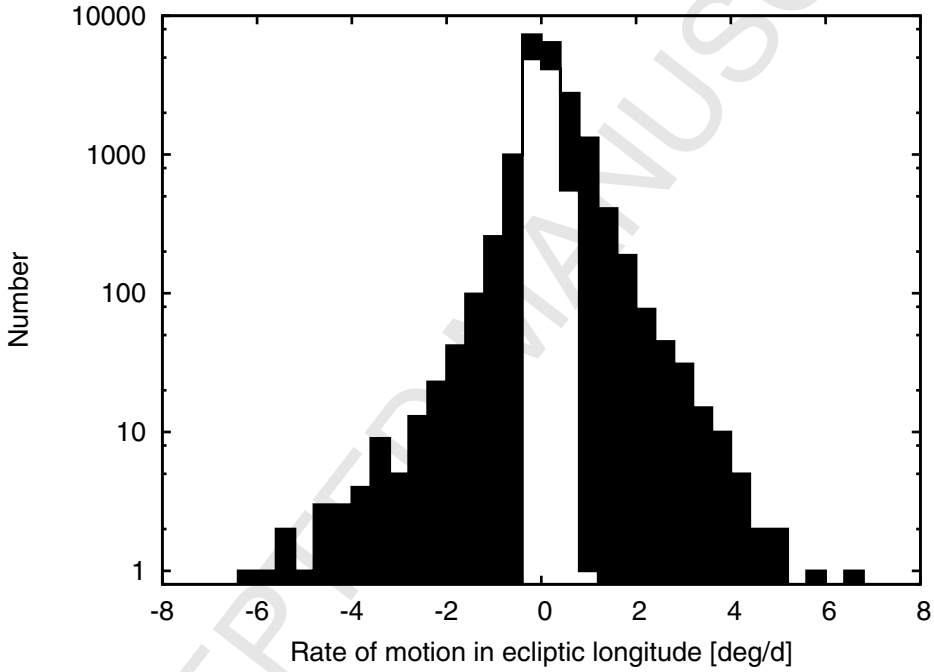



Figure 2:

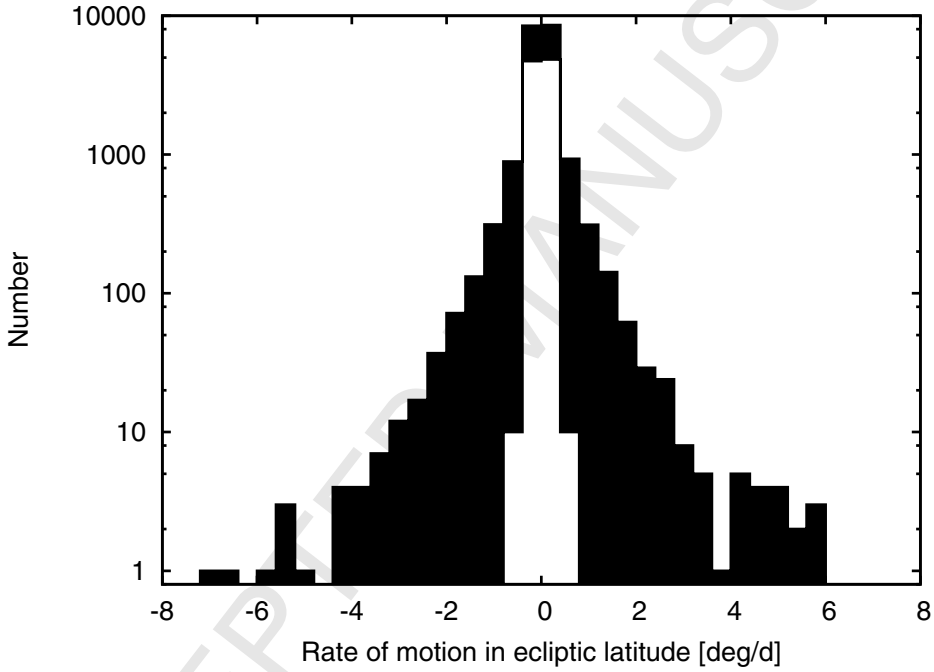



Figure 3:

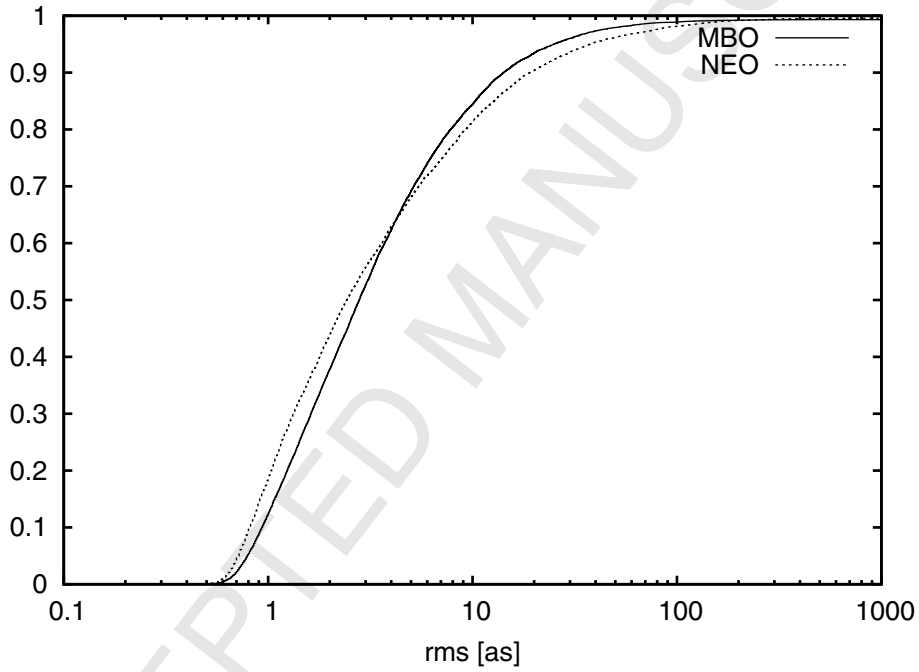



Figure 4: