StarGO: A New Method to Identify the Galactic Origins of Halo Stars

Zhen Yuan1,2,3, Jiang Chang4, Projwal Banerjee2,3, Jiaxin Han2,3,5, Xi Kang4, and M. C. Smith1

1 Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, People’s Republic of China; sala.yuan@gmail.com
2 Department of Astronomy, Shanghai Key Laboratory for Particle Physics and Cosmology, Shanghai Jiao Tong University, Shanghai 200240, People’s Republic of China
3 IFSA Collaborative Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, People’s Republic of China
4 Purple Mountain Observatory, the Partner Group of MPI für Astronomie, 2 West Beijing Road, Nanjing 210008, People’s Republic of China
5 Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan

Received 2017 November 22; revised 2018 June 14; accepted 2018 June 15; published 2018 August 7

Abstract

We develop a new method, Stars’ Galactic Origin (StarGO), to identify the galactic origins of halo stars using their kinematics. Our method is based on a self-organizing map (SOM), which is one of the most popular unsupervised learning algorithms. STARGO combines SOM with a novel adaptive group identification algorithm with essentially no free parameters. To evaluate our model, we build a synthetic stellar halo from mergers of nine satellites in the Milky Way. We construct the mock catalog by extracting a heliocentric volume of 10 kpc from our simulations and assigning expected observational uncertainties corresponding to bright stars from Gaia DR2 and LAMOST DR5. We compare the results from STARGO against those from a friends-of-friends-based method in the space of orbital energy and angular momentum. We show that STARGO is able to systematically identify more satellites and achieve higher number fraction of identified stars for most of the satellites within the extracted heliocentric volume. When applied to data from Gaia DR2, STARGO will enable us to reveal the origins of the inner stellar halo in unprecedented detail.

Key words: Galaxy: formation – Galaxy: halo – Galaxy: kinematics and dynamics – methods: data analysis

1. Introduction

According to the hierarchical structure formation theory, the Milky Way (MW) grows to its current size through frequent accretion and merger events. During these violent processes, satellite galaxies are tidally disrupted and the disk gets heated. The stellar halo is built up at the same time as a repository of stars from various origins (Bullock & Johnston 2005; Font et al. 2006; De Lucia & Helmi 2008; Cooper et al. 2010; Deason et al. 2016). Due to the approximately dissipationless nature of stars, substructures in the stellar halo, such as the stellar debris from a satellite or groups of stars that originated from the Galactic disk, may retain the memory of their origins. The identification of these substructures is the first step toward unraveling the evolution history of the MW. A number of such substructures have been found in the last decade, adding strong support to the scenario of hierarchical structure formation. One famous example is the discovery of the Sagittarius dwarf galaxy (Ibata et al. 1994, 1995; Yanny et al. 2000) and its tidal streams (Mateo et al. 1996; Ibata et al. 2001; Majewski et al. 2003), both of which are located in the stellar halo.

The current hierarchical structure formation paradigm implies that the inner stellar halo contains a wealth of information about the early assembly history of the MW, as the stars there tend to be accreted a long time ago. However, identifying substructures in configuration space is not easy due to the fact that the accreted substructures in the inner stellar halo have undergone mixing for a long time. Furthermore, this region is also populated by star groups likely originated from the disk, e.g., Monoceros (Bergemann et al. 2018; Laporte et al. 2018), which makes substructure identification from satellites difficult.

On the other hand, identifying substructures in phase space can be relatively easier, given the additional information from the velocities. In particular, the separations of stars in the integral-of-motion space are much better conserved, and thus provide a natural coordinate system for identifying the original grouping of stars (Helmi et al. 1999, 2006; Klement et al. 2009; Smith et al. 2009; Smith 2016). Previous searches of substructures in the inner stellar halo were hindered by the limited astrometric data. With the advent of Gaia, we now have 5D astrometric data for an unprecedented number of stars (1332 million) from Data Release 2 (Lindegren et al. 2018). Cross matching the Tycho-Gaia Astrometric Solution (TGAS; Gaia Collaboration et al. 2016) with other surveys such as RAVE (Casey et al. 2017), LAMOST (Luo et al. 2015), 2MASS (Skrutskie et al. 2006), and APOGEE (Anders et al. 2014) has produced a stellar library within ~20 kpc and has already led to several discoveries. For example, Koposov et al. (2017) discovered faint MW satellites by searching for overdensities in configuration space, Helmi et al. (2017) found a substructure of halo stars in integral-of-motion space, and Myeong et al. (2017) identified the existence of a comoving star cluster with additional information of metallicity distribution.

Despite the increasing discovery of identified substructures, their number is far below the predictions from ΛCDM cosmology. According to Aquarius simulations (Springel et al. 2008), hundreds of streams are expected in the Gaia sky (Gómez et al. 2013; Maffione et al. 2015). To systematically identify these substructures using the vast amount of astrometric data, several methods have been developed, including distance based methods such as friends-of-friends (FoF; Helmi & de Zeeuw 2000), and density-based algorithms.
such as Mean Shift (Gómez & Helmi 2010) and Watershed (Helmi et al. 2017).

In this paper, we propose a new method of substructure identification that requires essentially no free parameters. Our method utilizes a machine-learning technique called self-organizing map (SOM) (Kohonen 2001), that maps out the topology of a high-dimensional data set onto a 2D map. Using the fact that stars with the same origin have similar orbital energy and angular momentum, we first apply SOM to the n-Dimensional (n-D) input space constructed from these quantities and visualize the results in a 2D neural map. Then, we develop a new adaptive group identification scheme based on the resulting 2D map. As SOM retains the topological structure of the data set, it can manifest the fine structures in the data. This makes our method particularly well suited in identifying groups that are weakly clustered. We test the performance of our method by applying it to a mock catalog generated from our simulation of a MW-like system with realistic observational uncertainties.

The paper is organized as follows. Section 2 describes the details of model setup and simulations for generating the mock catalog; Section 3 discusses the details of SOM and group identification of STARGO; Section 4 presents the results of STARGO applied to the mock catalog and comparisons with FoF; and Section 5 provides our conclusion.

2. Simulations

2.1. Overview

A popular approach of building a stellar halo is using zoom-in ΛCDM cosmological simulation of a MW-like system accompanied by post-process of star tagging using semi-analytic models such as GALFORM (De Lucia & Helmi 2008; Cooper et al. 2010; Tumlinson 2010). Although such models can retain the realistic accretion history of a MW-like system, they do not include any stellar components, such as the disk and the bulge, which are crucial for modeling the kinematics of stars in the inner stellar halo.

Another approach involves using an analytic potential for the dark matter and stellar components of the MW, while using N-body models for the dark matter component of satellites (Helmi & de Zeeuw 2000). The main advantage of this approach is that it can achieve a higher resolution than pure N-body simulations. Instead of using the actual merger history, an artificial one is used to build up a synthetic stellar halo. Gómez et al. (2010) used a similar approach but with a time dependent MW potential. In such studies, stars in satellites are assigned to particles in post-process and in situ stars are added as background contamination. In a different approach of resimulation of a MW-like system, Jean-Baptiste et al. (2017) modeled the MW and satellites as collections of both dark matter and star particles to get a live N-body simulation.

In this study, we use a static analytic potential to model the dark matter halo of the MW, while particles are used to model stars in the disk and the bulge. For satellites, particles are used for both dark matter and stars. Although our method cannot account for dynamical friction since we use an analytic potential, it is expected to have a minor effect for the mass range of satellites chosen in our study (Amorisco 2017; Frings et al. 2017). This makes our method computationally much less expensive compared to studies which use live models (e.g., Jean-Baptiste et al. 2017).

Similar to the second approach mentioned above, we use an artificial merger history. Zoom-in cosmological simulations of MW-like systems suggest that the main contributors to the inner stellar halo are a few satellites that infall at early times (De Lucia & Helmi 2008; Cooper et al. 2010). In this study, we build up an synthetic stellar halo through several minor mergers following the same ideas as Boylan-Kolchin et al. (2008) and Amorisco (2017), where the dynamical friction from dark matter halo can be neglected. On the other hand, energy and angular momentum exchange between the MW center (the disk and the bulge) and satellites is important since all the satellites have pericenter distances within 20 kpc (see Table 2). This however, is automatically taken into account as they are modeled using particles. Here, we model the accretion of nine satellites which infall in the first 0–4 Gyr (see Table 2), where the total simulation time is 12 Gyr.

We use progenitors with infall mass ratios relative to the virial mass of MW of ≤1:25. In this mass range, even under dynamical friction from the halo, the initial orbital imprints of satellites can be retained in some of the stars (Amorisco 2017). The infall radial velocity, , and tangential velocity, , are set to be some fraction of the virial velocity of the MW, . The adopted value of the fractions are taken from the preferred range from cosmological simulations by Jiang et al. (2014). We calculate the circularity , defined as the ratio of the total angular momentum to the angular momentum for a circular orbit of the same energy. The satellites in our model have , which are consistent with values found in studies of accreted satellites from cosmological simulations (Jiang et al. 2014), as well as from other models of the stellar halo (Amorisco 2017).

2.2. Our Model

The dark matter halo of the MW is described as a Navarro-Frenk-White (NFW) potential (Navarro et al. 1997) with virial mass such that and concentration parameter . The stellar part of the MW consists of a Hernquist bulge (Hernquist 1990) and an exponential disk with the total stellar mass of , and particle mass of . The bulge component contributes ~20% of the total stellar mass, with a density profile given by

\[ \rho_b(r) = \frac{M_b}{2\pi r (r+a)^3}, \]

where , , and are the mass, scale length, and radius, respectively. The rest ~80% of the total stellar mass is in the stellar disk of mass with the density profile parameterized by scale length and scale height given by

\[ \rho_d(r,z) = \frac{M_d}{4\pi z_0 R_s^2} \text{sech}^2\left(\frac{z-z_0}{2z_0}\right)\exp\left(-\frac{R}{R_s}\right), \]

where and are the height and radius, respectively.

For pre-cooked progenitors of satellites, we use a similar recipe to Chang et al. (2013). Each satellite has a dark matter halo with an NFW profile, and an exponential disk with a total stellar mass of 1% of the virial mass of the satellite. We use particle mass of and for the dark matter halo and stellar disk, respectively. We design three types of progenitors: H-m, M-m, and L-m, with total masses of and , respectively (see Table 1).
Using H-m, M-m, and L-m, we create nine satellites (sat1–9) with distinct infall scenarios by changing their initial velocities and positions (see Table 2). All of the satellites are released at distance $R_{\text{ini}}$ with an initial radial velocity $V_r$ and tangential velocity $V_\theta$ (Benson 2005; Jiang et al. 2014). The inclination of the satellite orbit with respect to the disk is characterized by $i$, which is the angle between the initial orbital direction of the satellite and the initial Galactic $z$ direction. The values of $i$ are chosen from a range of $0^\circ$–$60^\circ$. We choose $R_{\text{ini}}$ to be less than the virial radius for all the satellites to emulate the early infall scenarios when the MW is smaller.

### 2.3. Catalog

To generate data sample from our simulation, we select all the stars within 10 kpc relative to the Sun, which is taken to be at the galactocentric distance $R_g = 8$ kpc. This is done to get data with the coverage similar to the Gaia sky. We address the sampling bias introduced by the solar position by generating eight samples, where the solar position is rotated by $45^\circ$ each time in the $x$–$y$ plane. We find the samples to be quantitatively similar, i.e., number of stars in each population is similar. In this paper, we use only one of the samples for demonstration, the details of which are given in Table 2.

The disk and the bulge population accounts for more than 95% in our sample, most of which are distributed close to the Galactic disk plane. As we are mainly interested in the halo stars, we exclude all the stars from these two components to generate the mock catalogues. In real observational samples, this can also be done relatively easily by using a cut for metallicity or distance from the mid-plane. The total number of stars in each satellite is denoted by $N_{\text{sat}}$ (see Table 2). Most of the stars in the extracted heliocentric volume come from six of the nine satellites (sat1, sat2, sat4, sat5, sat7, and sat9).

K-giants are ideal tracers of the stellar halo because they are bright, and distances can be reliably estimated from photometry. Therefore, we construct a mock catalog of K-giant stars, adopting errors similar to what will be obtained for a sample of such stars from LAMOST and Gaia DR2. Specifically, we assign the distance error to be 20% according to the distance estimation method using photometry (Xue et al. 2014). We adopt the error of radial velocity to be $7$ km s$^{-1}$, which is consistent with LAMOST (Schönrich & Aumer 2017). The proper motion error at $G \sim 16$–$17$ mag is $0.1$–$0.2$ mas yr$^{-1}$ from Gaia DR2 (Lindegren et al. 2018). We take $0.15$ mas yr$^{-1}$ for all the stars in the catalog. This is a reasonable number for K-giants, as even at 10 kpc, they have $G \sim 16$–$17$ mag. The corresponding error in tangential velocity for a star at 10 kpc is about $7$ km s$^{-1}$, i.e., comparable to our assumed error in radial velocity from low-resolution spectroscopy.

Before we discuss the details of our group identification method and the results, it is useful to first visualize the input data. We plot stars in the Aitoff projection of angular momentum space shown in Figure 1. The left panel shows stars from different satellites, whereas the right panel shows the corresponding density map. We can clearly see that stars of the same origin tend to cluster, although stars of different origins often overlap each other. This can also be seen from Figures 2 (a)–(c), where stars are plotted in $(L_x, L_y)$, $(L_x, L_z)$, and $(L_x, E)$. The clustering can be clearly seen in the corresponding density maps (see the right panel of Figures 1 and 2(d)–(f)). This clustering information is the basis of all substructure identification methods. Below, we discuss the details of our method applied to different input spaces discussed above.

### Table 1

Properties of the MW and Satellite Galaxies

| Type | $M_{\odot}$ (M$_\odot$) | $V_{\text{vir}}$ (km s$^{-1}$) | $c$ | $N_b$ | $M_{\odot}$ | $R_{\text{vir}}$ (kpc) | $N_{\text{sat}}$ | $m_{\text{p, sat}}$ (M$_\odot$) |
|------|-----------------|----------------|-----|------|-------------|------------------|-------------|-----------------|
| MW   | $10^{12}$       | 162.6          | 7.0 |      | $3 \times 10^6$ | 3.01             | 6 $\times 10^3$ | 5 $\times 10^3$ |
| H-m  | $4 \times 10^{10}$ | 55.6           | 9.3 | $7.84 \times 10^3$ | 8 $\times 10^3$ | 0.98             | 1.6 $\times 10^3$ | 5 $\times 10^3$ |
| M-m  | $10^{10}$       | 35.0           | 10.6| $1.96 \times 10^3$ | 2 $\times 10^3$ | 0.57             | 4 $\times 10^3$ | 5 $\times 10^3$ |
| L-m  | $2.5 \times 10^9$ | 22.1           | 12.2| $4.9 \times 10^3$ | $5 \times 10^3$ | 0.34             | $10^3$        | 5 $\times 10^3$ |

Note. For each type of galaxy, $M_{\odot}$ and $V_{\text{vir}}$ are the viral mass and viral velocity (for satellites these values refer to their initial values), respectively. Note that $R_{\text{vir}}$ has the same value as $V_{\text{vir}}$ with the unit of kpc/h. For the dark matter halo part, $c$ is the concentration parameter of the NFW model, $N_b$ is the number of particles in the dark matter halo, and $m_{\text{p, sat}} = 5 \times 10^4$ M$_\odot$ is the mass for each dark matter particle. For the stellar part, $M_{\odot}$ is the stellar mass, $R_{\text{vir}}$ is the disk scale length, $N_{\text{sat}}$ is the number of star particles, and $m_{\text{p, sat}}$ is the mass for each star particle.

### Table 2

Details of the Mock Catalog

| Type | $T_{\text{ini}}$ (Gyr) | $R_{\text{ini}}/R_{\text{vir,MW}}$ | $(V_r, V_\theta)/(V_r, V_\theta, \text{MW})$ | $i$ (°) | $r_{\text{peri}}$ (kpc) | $j = J/J_{\text{vir}}$ | $N_{\text{sat}}$ |
|------|-----------------|----------------|----------------|------|------------------|-------------|-------------|
| sat1 | H-m             | 0              | 0.4            | (0.64, -0.64) | 60    | 22.8             | 0.71        | 3,609       |
| sat2 | M-m             | 0              | 0.4            | (0.96, 0.32) | 45    | 9.1              | 0.32        | 1,970       |
| sat3 | L-m             | 0              | 0.4            | (1.0, 0.2)   | 30    | 5.4              | 0.20        | 59          |
| sat4 | H-m             | 2              | 0.6            | (0.96, 0.32) | 45    | 13.2             | 0.32        | 5,598       |
| sat5 | M-m             | 2              | 0.6            | (0.96, -0.32)| 0     | 13.2             | 0.32        | 1,961       |
| sat6 | L-m             | 2              | 0.6            | (0.72, 0.72) | 45    | 38.7             | 0.71        | 9           |
| sat7 | H-m             | 4              | 0.8            | (0.6, -0.2)  | 0     | 10.6             | 0.32        | 4,985       |
| sat8 | M-m             | 4              | 0.8            | (0.36, 0.36) | 30    | 22.2             | 0.71        | 38          |
| sat9 | L-m             | 4              | 0.8            | (0.48, 0.16) | 15    | 8.4              | 0.32        | 341         |

Note. The initial condition of each satellite before it falls into the MW at $T_{\text{ini}}$ are represented by the distance $R_{\text{ini}}$, velocity $(V_r, V_\theta)$, and inclination angle $i$. The percenter distance and circularity of the orbit for each satellite are denoted as $r_{\text{peri}}$ and $j$, respectively. The number of stars from each satellite $N_{\text{sat}}$ is shown in the last column.
3. Method

Here, we develop a substructure identification method based on SOM, which belongs to unsupervised learning domain. We apply SOM to the mock data catalog, followed by a novel group identification procedure that utilizes the visualization of SOM output in the 2D neural map. Below, we give a brief

Figure 1. Aitoff projection of angular momentum of stars from mock catalog in the galactic reference frame. Left panel: stars from different satellites are assigned unique colors as follows: sat1 (salmon), sat2 (olive), sat3 (magenta), sat4 (purple), sat5 (pink), sat6 (blue), sat7 (cyan), sat8(gold), and sat9(light blue). Right panel: the corresponding density map of all the satellite stars in the same projected space as in the left panel.

Figure 2. Top row: projection of all the stars from the mock catalog in (a) $L_x$, (b) $L_x-L_z$, and (c) $L_z$ plane, using the same color codes of Figure 1. Bottom row (b)–(d): the corresponding density map of all the stars in the same plane as the top row.

Figure 3. Results from the first iteration of workflow of STARGO to the mock catalog in the $(E, L_x, L_y, L_z)$ space. (a) The blue shaded histogram shows the distribution of $u$, where the cyan and salmon dashed lines denote $u_m$ and $u_{thr}$ values, respectively. (b) 2D neural map resulting from SOM, where the $u$ value between adjacent neurons is represented by the gray color scale. (c) The same map as (b) which shows the selected neurons in different colors according to step 2–3 of the workflow. The neurons with $u < u_m$ are marked by colored pixels; each pixel associated to a group is colored cyan, while groups with more than 30 stars have their pixels enclosed with a blue box. (d) The neurons with $u < u_{thr}$ are marked by salmon pixels according to step 4 of the workflow.
The Astrophysical Journal, 863:26 (11pp), 2018 August 10

Yuan et al.

![Diagram](image)

Figure 4. Workflow of STARGO.

introduction to SOM followed by details of our novel group identification procedure.

### 3.1. Self-organizing Map

The aim of SOM is to map a n-D input data to a 2D neural map, while retaining the topological structures within the data at the same time. The starting point is the construction of a 2D grid of neurons with the same dimension and range as the n-D input vectors \( v \). Given an input vector \( v^i \), for the \( i \)th star from the data catalog, we first find the neuron that has the closest weight vector to \( v^i \) by finding the neuron with the minimum value of \( |w - v^i| \). Such a neuron is defined to be the best matching unit (BMU). The learning process involves improving weight vectors \( w \) of all the neurons toward \( v^i \) according to their distances \( d_{ab}^{i,b} \) to the BMU located at \((a, b)\) on the 2D neural map, where \( d_{ab}^{i,b} \) is defined as

\[
\begin{align*}
  d_{ab}^{i,b} &= \sqrt{(a - a^i)^2 + (b - b^i)^2}. \\
  \text{(3)}
\end{align*}
\]

The change in the weight vector \( dw_{ab}^{i,b} \) due to the \( i \)th star is given by

\[
\begin{align*}
  dw_{ab}^{i,b} &= \alpha_q \exp \left( \frac{-d_{ab}^{i,b}}{\sigma_q^2} \right) (v^i - w_{ab}^{i,b}), \\
  \text{(4)}
\end{align*}
\]

where \( \alpha_q \) characterizes the learning rate, and \( \sigma_q \) controls the neighboring influence of neurons around the BMU for the \( q \)th iteration. We can see from the above equation that the change in the weight of a neuron is sensitive to its distance from the BMU. The learning process is performed using \( v^i \) for each star in the data set. The learning process is then repeated for a total number of \( N_{\text{iter}} \) iterations. For the \( q \)th iteration, the corresponding \( \alpha_q \) and \( \sigma_q \) is

\[
\begin{align*}
  \alpha_q &= \alpha_0 (1 - q/N_{\text{iter}}), \\
  \sigma_q &= \sigma_0 (1 - q/N_{\text{iter}}), \\
  \text{(5)}
\end{align*}
\]

where we use typical fiducial values of \( \alpha_0 = 0.3 \) and \( \sigma_0 = \max(m, n)/2 \) similar to Geach (2012). We find that the results are independent of \( \alpha_q \) and \( \sigma_q \) for reasonable variation around this fiducial value. As the number of iteration \( q \) increases, \(|dw|\) is reduced due to the decrease in \( \alpha_q \) and \( \sigma_q \), leading to refinement of the learning process. The learning process is considered to be complete when \(|dw|/|w| \to 0\).

#### 3.2. Group Identification

We feed the mock catalog in a given input space to a 80 \( \times \) 80 neural network. After the application of SOM, the clustering structures can be visualized by using the differences between the weight vectors of neighboring neurons, which are the elements of the \( u \)-matrix defined as

\[
\begin{align*}
  u_{ab} &= \log_2(|w_{ab+1} - w_{ab}| + |w_{a,b+1} - w_{a,b}|). \\
  \text{(6)}
\end{align*}
\]

Neurons mapped to stars in highly clustered regions tend to have similar angular momentum and orbital energy, leading to lower values of \( u \) and vice versa. Figure 3(a) shows the distribution of \( u \) while Figure 3(b) shows the resulting 2D 80 \( \times \) 80 neural map, where \( u \) is represented by the gray color scale. Each star can be mapped to its BMU in the 2D map. We note that every neuron can be associated with more than one star or no stars at all.

Based on the 2D map generated by SOM, we develop a novel algorithm for identification of substructures. Below, we list the steps adopted for group identification starting with the application of SOM followed by our algorithm.

1. We first normalize each component of the input vector.
   For each dimension of the input vector, we calculate the 95% confidence interval for the whole sample. We then divide each component of the input vector for each star by this normalizing factor. We then apply SOM to all the
Figure 5. Illustration of STARGO workflow applied to the mock catalog in the \((E, I_X, I_Y, I_Z)\) space for three iterations. (Ia) Direct map of stars to their BMUs after the initial iteration of SOM. Stars in each satellite are assigned with unique color (same as Figure 1) and symbol: sat1 (salmon upper triangle), sat2 (olive right triangle), sat3 (magenta plus), sat4 (purple diamond), sat5 (pink cross), sat6 (blue square), sat7 (cyan star), sat8 (gold hexagon), and sat9 (light blue left triangle). (Ib) On the same map as (Ia), the neurons in seed groups are marked by blue pixels, same as Figure 3(c). The neurons with \(u < u_{th} \) are marked by salmon pixels, same as Figure 3(d). (Ic) On the same map as (Ia), the identified star groups (Group A–G) are plotted, where stars in each group are mapped to their BMU with the same color coding as (Ia). (Ila) Direct map of stars from Group A after the second iteration of SOM. (Ilb) On the same map as (Ila), the selected neurons according to step 3–4 from the second iteration of the workflow are plotted. (IIC) On the same map as (Ila), the identified star groups (Group A0–A6) are plotted. (IIa) Direct map of stars from Group A3 after the third iteration of SOM. (IIb) On the same map as (IIa), the selected neurons are plotted from the third iteration of the workflow. (IIc) On the same map as (IIa), the identified star groups (Group A3a–A3b) are plotted.
stars in the normalized input space and calculate \( u \)-matrix from the resulting map. We associate each star to its BMU in the 2D map.

2. We group neighboring neurons with \( u < u_m \) to form candidate seed groups (marked by cyan pixels in Figure 3(c)), where \( u_m \) is the median value of \( u \) for all neurons in the map.

3. A candidate seed group resulting from Step 2 is considered as a bona fide seed group (enclosed with blue boxes in Figure 3(c)) if more than 30 stars are associated to neurons in the group.

4. If there are more than one bona fide seed group, we maximize the size of each group by increasing the value of \( u_{thr} \) (shown as salmon dashed line in Figure 3(a)). This results in the merging of multiple groups and we increase \( u_{thr} \) until we have two groups remaining from our original set of seed groups (marked by salmon pixels in Figure 3(d)). We stop increasing \( u_{thr} \) when these two groups are as large as possible, i.e., if \( u_{thr} \) was increased any further then these would merge. Additional groups can arise as the increased \( u_{thr} \) results in the formation of some new groups that have more than 30 stars associated to them.

5. Star associated with the identified neuron groups form the corresponding identified star groups. For each identified group, we apply SOM followed by the above group identification procedure by repeating steps 1 to 4 (see workflow in Figure 4). Stars not belonging to any group are designated as “unidentified.”

6. We stop the group identification procedure when no more than one seed group can be found after step 3.

The group identification algorithm above allows us to find substructures adaptively for a given data set. At each iteration, neurons with \( u < u_m \) correspond to stars that have clustering above the median value. We set the minimum number of stars for seed groups to be 30 to discard small groups with low significance. The threshold of identified group is also set to be 30, which is slightly below the number of stars from the smallest population in the catalog. Increasing the value of \( u_{thr} \) in step 4 is designed to maximize the completeness of grouped stars. Because we apply our algorithm iteratively to each group, the final set of groups represent the smallest indivisible group that has at least 30 stars.

### 3.3. Size and Convergence Check

We check the dependence of the results on the size of neural network by using networks of size \( 50 \times 50, 80 \times 80, \) and \( 100 \times 100 \) neurons. We find that the convergence is achieved for network size of \( 80 \times 80 \), with the larger network yielding almost identical results. The smaller network fails to identify fine structures due to coarse gridding. Thus, we use the network of size \( 80 \times 80 \) throughout the study. We perform additional checks on dependence of the results on the iteration number \( N_{iter} \). We use \( N_{iter} = 200, 300, \) and 400, finding the results are already converged for \( N_{iter} = 200 \). We adopt \( N_{iter} = 200 \) as the default value.

### 4. Results and Discussions

In this section, we apply STARGO to the mock catalog in the \((E, L_x, L_y, L_z)\) space. We compare the results with the corresponding results using FoF. We know that in axisymmetrical potential of the MW, the orbital parameters \( E, L_x, \) and \( L_y \) are known to be approximately conserved, whereas \( L_y \) and \( L_z \) evolve coherently (Helmi & de Zeeuw 2000; Knebe et al. 2005; Klement 2010; Gómez & Helmi 2010; Maffione et al. 2015). In a more realistic scenario, such as in our model which includes a live N-body disk and bulge, the conservation of these quantities are more strongly violated. However, after stars from a satellite interact with the stars in the MW and other satellites, some of them can still have similar \( E \) and \( L_z \), due to very similar disruption history. This can be seen from the Aitoff projection of L shown in \((\theta, \phi)\) in the left panel of Figure 1, where stars of the same origin tend to cluster. Similar clustering can also be seen in \((L_x, L_y), (L_x, L_z)\), and \((L_y, E)\) (see Figures 2(a)–(c)). Substructure identification methods can exploit these clusterings to identify star groups.

#### 4.1. STARGO

Following steps 1–6 of the workflow (see Figure 4), we apply SOM to the mock catalog in the \((E, L_x, L_y, L_z)\) space. Figure 5(a) shows the training results after the application of SOM (step 1 of the algorithm) on the 2D neural map. Each BMU is represented with the color and symbol according to the
Table 3
Results from STARGO Applied to the Mock Catalog in the \((E, L_x, L_y, L_z)\) Space

| Group ID | A   | B   | C   | D   | E   | F   | G   |
|----------|-----|-----|-----|-----|-----|-----|-----|
| i=0      |     |     |     |     |     |     |     |
| Sat1     | 63  | 35  |     |     | 50  | 61  |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) | 48  | 39  | 63  | 100 | 50  | 66  |     |
| i=1      |     |     |     |     |     |     |     |
| GrpID    |     |     |     |     |     |     |     |
| Sat1     | 48  | 39  | 63  | 100 | 50  | 66  |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) | 54  | 90  | 100 |     | 100 | 92  |     |
| i=2      |     |     |     |     |     |     |     |
| GrpID    | A0  | A1  | A2  | A3  | A4  | A5  | A6  |
| Sat1     | 180 | 35  | 52  |     |     |     |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) | 33  | 86  | 52  |     | 44  | 74  | 64  |
| i=3      |     |     |     |     |     |     |     |
| GrpID    | A0a | A0b | A3a | A3b |     |     |     |
| Sat1     |     |     |     |     |     |     |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) | 208 | 1887 | 284 |     |     |     |     |
| i=4      |     |     |     |     |     |     |     |
| GrpID    |     |     |     |     |     |     |     |
| Sat1     | 131 | 93  | 80  |     |     |     |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) |     |     |     |     |     |     |     |
| i=5      |     |     |     |     |     |     |     |
| GrpID    | A0a |     |     |     |     |     |     |
| Sat1     |     |     |     |     |     |     |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) |     |     |     |     |     |     |     |
| i=6      |     |     |     |     |     |     |     |
| GrpID    | A0a |     |     |     |     |     |     |
| Sat1     |     |     |     |     |     |     |     |
| Sat2     |     |     |     |     |     |     |     |
| Sat3     |     |     |     |     |     |     |     |
| \(N_{grp}\) purity\%(iteration) |     |     |     |     |     |     |     |

**Note.** For each iteration, we list the identified satellites and the number of stars in the corresponding identified groups. \(N_{grp}\) denotes the total number of stars in each identified group. Groups that require further iteration are highlighted in bold at each step.
satellite of the associated stars, where the darker symbols represent multiple stars mapped to the same neuron. Some neurons are BMUs of stars belonging to different satellites, which can be seen with overlaid symbols. After we perform the group identification algorithm (steps 2–4), the neurons in seed groups are marked by blue pixels and the neurons with $u < u_{thr}$ are marked by salmon pixels in Figure 5(b), same as Figures 3(c)–(d). Figure 5(c) shows the identified star groups.

Table 4

| GrpID | A | B | C | D | E | F | G | H | I | J | K |
|-------|---|---|---|---|---|---|---|---|---|---|---|
| Sat1  |   |   |   |   | 41|   |   |   |   |   |   |
| Sat4  |   |   |   |   | 29|   | 35| 22| 33|   |   |
| Sat5  | 102|   |   |   |   |   |   |   |   |   |   |
| Sat7  |   | 47| 42|   |   |   |   |   |   |   |   |

$N_{grp}$ purity (%)

- $\eta = 0.15$
  - $\eta = 0.20$
  - $\eta = 0.25$

Table 5

| GrpID | A | B | C | D | E | F | G | H | I | J | K | L | M |
|-------|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sat1  | 272| 95| 72| 67| 57| 43| 40| 36| 31|   | 28|   | 31|
| Sat2  |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Sat4  |   |   |   |   |   |   |   |   |   |   |   |   |   |
| Sat9  |   |   |   |   |   |   |   |   |   | 67|   |   |   |

$N_{grp}$ purity (%)

- $\eta = 0.15$
  - $\eta = 0.20$
  - $\eta = 0.25$

satellite of the associated stars, where the darker symbols represent multiple stars mapped to the same neuron. Some neurons are BMUs of stars belonging to different satellites, which can be seen with overlaid symbols. After we perform the group identification algorithm (steps 2–4), the neurons in seed groups are marked by blue pixels and the neurons with $u < u_{thr}$ are marked by salmon pixels in Figure 5(b), same as Figures 3(c)–(d). Figure 5(c) shows the identified star groups.
(Group A–G) using the same color coding as Figure 5(a). The group identification applied to individual groups requires six more iterations for Group A and one more iteration for Group E before we reach the end of the workflow (see Figure 4). For illustration, we show the detailed group identification procedure for two more iterations for Group A in Figures 5(II)–(III).

Figure 6 shows the full schematic diagram for the hierarchical group identification from STARGO with the detailed results listed in Table 3.

We find that for most identified groups, the major contribution is from a single satellite. If the purity for a group is \( \geq 60\% \), we identify the group with the corresponding satellite, which we refer to as the dominant contributor. Using this criteria, STARGO is able to find a total of 24 star groups, out of which 21 can be identified with satellites. One group is considered as a spurious group (Group A3a), which has roughly equal fraction of stars from sat4 and sat7 with purity \( \sim 40\% \). The remaining two groups (Group B and A2) have purity ranging from 50\%–54\% and thus cannot be strictly identified with a satellite using our criteria. We find that all of the six major satellites (sat1, sat2, sat4, sat5, sat7, and sat9) in the mock catalog can be identified with at least one group. The number fraction \( f \) of stars of a satellite in the extracted volume that are identified with its corresponding groups is \( \geq 5\% \) for all major satellites except sat2 (\( f = 1.7\% \)). This is likely due to the fact that sat2 is more heavily disrupted compared to the other five major satellites. On the other hand, for the two largest contributors to the mock catalog, sat4 and sat7, STARGO is able to identify 6.6\% (four groups) and 12\% (seven groups) of the stars, respectively. Overall, STARGO is able to identify a total of 1850 stars from satellites within the analyzed volume that are the major contributors to the identified groups. This constitutes a fraction \( f_{\text{tot}} = 10\% \) of the total number of stars in the mock catalog.

4.2. Friends-of-friends

A widely used method of substructure identification is FoF, which is the standard procedure of finding halo groups used in cosmological simulations. In this case, the linking length \( L_p \) is a key parameter, which is chosen empirically. The typical value of \( L_p \) is set to be 0.2 times the interparticle distance, which is a characteristic length scale used in the definition of dark matter halo. When FoF is applied to substructure identification in the integral-of-motion space, all “distances” have units of angular momentum or energy, such that interparticle distances lose their physical meaning. Following (Helmi & de Zeeuw 2000), we set the characteristic length scale to be the dispersion of the total angular momentum of stars \( \Delta L_{\text{cluster}} \). We note that, similar to STARGO, we use a normalized input space which is dimensionless. Thus, the dispersion of the angular momentum can be used as a scale for every dimension. The linking length \( L_p \) is set to be \( \eta \Delta L_{\text{cluster}} \), where \( \eta \) is empirically chosen from a range of \( \pm 0.1–0.2 \). Finding optimal values of \( \eta \) and determining \( \Delta L_{\text{cluster}} \) are problematic for substructure identification.

Following the procedures from Helmi & de Zeeuw (2000), we test the performance of FoF applied to the mock catalog. We estimate \( \Delta L_{\text{cluster}} \) from the distribution of \( L_p \). Specifically, we set it to be half of 68\% confidence interval, which roughly corresponds to the ±1\( \sigma \) range for a normal distribution. We apply FoF using different values of \( \eta \). In order to visualize the results from FoF for easy comparisons with STARGO, we again use the 2D neural map. To do this, we map each star to its BMU resulting from the initial application of SOM. We use distinct colors to plot the stars of each group identified by FoF. As in the case of STARGO, we only consider groups with more than 30 stars. Figures 7(a)–(e) shows the results for three different values of \( \eta \). As we can see, the group identification is sensitive to \( \eta \). The optimal results are found for \( \eta = 0.15–0.25 \), which gives the maximum number of identified satellites within the extracted heliocentric volume and highest \( f_{\text{tot}} \). The results for \( \eta = 0.15, 0.20, \) and 0.25 are listed in Table 4.

For \( \eta = 0.20 \), FoF is able to identify sat1, sat4, sat5, sat7, and sat9 with \( f_{\text{tot}} = 3.9\% \), where six out of seven groups can be identified with satellites. When \( \eta \) is reduced to 0.15, FoF is still able to identify sat1, sat4, sat5, sat7 with \( f_{\text{tot}} = 2.1\% \), but is unable to identify sat9. In this case, nine out of the 11 groups can be identified with satellites. On the other hand, for \( \eta = 0.25 \), FoF can identify sat1, sat2, sat4, sat9 with \( f_{\text{tot}} = 4.5\% \), where 12 out of 13 groups can be identified with satellites. In contrast, STARGO is able to identify all the satellites with \( f_{\text{tot}} = 10\% \). Interestingly, for all the three values of \( \eta \), the group that cannot be identified with any single satellite (marked by gray pixels in Figure 7) is the largest group, with roughly equal contributions from sat4 and sat7. This is likely due to the fact that sat4 overlaps heavily with sat7 (see Figures 1 and 2). It results in weak clustering features such that FoF is barely able to distinguish them. This can clearly seen from Table 5, where FoF gives very low values of \( f \) for sat4 and sat7 for \( \eta = 0.20 \) and is unable to identify sat7 for \( \eta = 0.25 \). \( \eta = 0.15 \) gives highest values of \( f = 2.1\%, 1.8\% \) for sat4 and sat7, respectively, which are still below \( f = 6.6\%, 12\% \) obtained from STARGO. Similarly, FoF also fails to identify sat2 for \( \eta = 0.15 \) and \( \eta = 0.2 \), and gives low value of \( f = 1.4\% \) for \( \eta = 0.25 \), whereas STARGO gives slightly better result of \( f = 1.7\% \). As mentioned before, this is likely due to the fact that sat2 has gone through severe disruption which results in weak clustering signal in the input space. Even for the optimal range of values of \( \eta \), the variation of FoF results can be seen from the fact that sat5 can be easily identified for \( \eta = 0.15 \) and \( \eta = 0.20 \) but cannot be identified at all for \( \eta = 0.25 \). On the other hand, STARGO gives higher value of \( f = 16\% \) compared to \( f = 14\% \) from the best case of FoF with \( \eta = 0.2 \). Similarly, the identified fraction of stars from satl increases sharply from 3.7\% to 20\% as \( \eta \) increases from 0.20 to 0.25. Compared to such variations, STARGO is able to identify sat1 with a moderate value of \( f = 12\% \).

For values of \( \eta \) outside of the optimal range, FoF identifies fewer satellites or has even lower values of \( f \) and \( f_{\text{tot}} \) within the analyzed volume (shown in Figure 7). For \( \eta = 0.1 \), FoF can find only one group of 36 stars (\( f_{\text{tot}} = 0.19\% \), which is associated with sat7. For \( \eta = 0.3 \), there are three groups identified from sat1, sat4 and sat9 with \( f_{\text{tot}} = 1.5\% \).

5. Conclusion

In this paper, we present a new substructure identification method STARGO that identifies and visualizes star groups hierarchically on top of a 2D neuron map. Our algorithm first maps the multidimensional phase space coordinates of stars into a 2D map using SOM while conserving the topological structure of the data set. It then identifies a hierarchy of star groups adaptively according to the significance of clustering at each step.

We test our algorithm using a mock catalog of stars within a heliocentric radius of 10 kpc generated from a simulated MW-like system, and compare the results against that from an FoF.
algorithm. In the tests we take into account observational errors that are expected for K-giants in the Gaia DR2 and LAMOST DR5 catalogs. In comparison to FoF, STARGO is able to identify star groups dominated by each of the six major satellites, whereas FoF is able to identify at most five, even after optimizing the linking length. In addition, STARGO can identify a higher fraction of stars from almost all the satellites compared to FoF (see Table 5). If we consider the number of stars from the dominant satellite in each group, we find that STARGO is able to identify a total of 10% of the total satellite population in the extracted heliocentric volume, whereas for FoF this fraction is below 4.5%.

In conclusion, STARGO is able to identify star groups efficiently by combining the sensitivity and visualization ability of SOM with an adaptive clustering algorithm. The adaptive group identification procedure allows us to systematically search for substructures while avoiding uncertainties from nuisance parameters. Overall, the results from STARGO are better than the results from FoF, even when using an optimized linking length. STARGO is an ideal tool to explore high-dimensional data set from the recently released Gaia DR2. Our method will be particularly useful for studies of the inner stellar halo, for example when applied to the cross-match of Gaia DR2 and spectroscopic surveys.

Z.Y. gratefully acknowledges Jingying Lin for inspiring discussions about SOM. Z.Y. thanks Xiangxiang Xue for sharing her expertise in applying FoF to the real data, and Chao Liu for insightful discussions about the algorithm of StarGO. Z.Y. is also indebted to Yi Peng Jing, Yu Luo, and Hong Guo for commenting on the early draft of this paper. All the authors thank the anonymous referee for valuable and constructive comments which greatly improved this work. This work is supported by the National Key Basic Research Program of China (No. 2015CB857003). Z.Y. and P.B. acknowledge the support of NSFC-11533006. J.C. and X.K. acknowledge the support of NSFC-11333008. J.X.H. acknowledge the support of Key Laboratory of Computational Astrophysics, National Astronomical Observatories, Chinese Academy of Sciences. M.C.S. acknowledges financial support from the CAS One Hundred Talent Fund and from NSFC grants 11637083 and 11333003. This work was also supported by the National Key Basic Research Program of China 2014CB845700.

We gratefully acknowledge the use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Shanghai Astronomical Observatory.

ORCID iDs

Zhen Yuan @ https://orcid.org/0000-0002-8129-5415
Projwall Banerjee @ https://orcid.org/0000-0002-6389-2697

References

Amorisco, N. C. 2017, MNRAS, 464, 2882
Anders, F., Chiappini, C., Santiago, B. X., & Rocha-Pinto, H. J. 2014, A&A, 564, A115
Benson, A. J. 2005, MNRAS, 358, 551
Bergemann, M., Sesar, B., Cohen, J. G., et al. 2018, Natur, 555, 334
Boylan-Kolchin, M., Ma, C.-P., & Quataert, E. 2008, MNRAS, 383, 93
Bullock, J. S., & Johnston, K. V. 2005, ApJ, 635, 931
Casey, A. R., Hawkins, K., Hogg, D. W., et al. 2017, ApJ, 840, 59
Chang, J., Macciò, A. V., & Kang, X. 2013, MNRAS, 431, 3533
Cooper, A. P., Cole, S., Frenk, C. S., et al. 2010, MNRAS, 406, 744
De Lucia, G., & Helmi, A. 2008, MNRAS, 391, 14
Deason, A. J., Mao, Y.-Y., & Wechsler, R. H. 2016, ApJ, 821, 8
Font, A. S., Johnston, K. V., Bullock, J. S., & Robertson, B. E. 2006, ApJ, 638, 585
Frings, J., Macciò, A., Buck, T., et al. 2017, MNRAS, 472, 3378
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, A&A, 595, A2
Geach, J. E. 2012, MNRAS, 419, 2633
Gómez, F. A., & Helmi, A. 2010, MNRAS, 401, 2285
Gómez, F. A., Helmi, A., Brown, A. G. A., & Li, Y.-s. 2010, MNRAS, 408, 935
Gómez, F. A., Helmi, A., Cooper, A. P., et al. 2013, MNRAS, 436, 3602
Helmi, A., & de Zeeuw, P. T. 2000, MNRAS, 319, 657
Helmi, A., Navarro, J. F., Nordström, B., et al. 2006, MNRAS, 365, 1309
Helmi, A., Veljanoski, J., Breddels, M. A., Tian, H., & Sales, L. V. 2017, A&A, 598, A58
Helmi, A., White, S. D. M., de Zeeuw, P. T., & Zhao, H. 1999, Natur, 402, 53
Hernquist, L. 1990, ApJ, 356, 359
Ibata, R., Irwin, M., Lewis, G. F., & Stolte, A. 2001, ApJL, 547, L133
Ibata, R. A., Gilmore, G., & Irwin, M. J. 1994, Natur, 370, 194
Ibata, R. A., Gilmore, G., & Irwin, M. J. 1995, MNRAS, 277, 781
Jean-Baptiste, I., Di Matteo, P., Haywood, M., et al. 2017, A&A, 604, A106
Jiang, J., Helly, J. C., Cole, S., & Frenk, C. S. 2014, MNRAS, 440, 2115
Klement, R., Rix, H.-W., Flynn, C., et al. 2009, ApJ, 698, 865
Klement, R. J. 2010, A&ARv, 18, 567
Knebe, A., Gill, S. P. D., Kawata, D., & Gibson, B. K. 2005, MNRAS, 357, L35
Kohono, T. 2001, Self-organizing Maps (3rd ed.; Berlin: Springer), 30
Koposov, S. E., Belokurov, V., & Torrealba, G. 2017, MNRAS, 470, 2702
Laporte, C. F. P., Gómez, F. A., Helmi, A., et al. 2013, MNRAS, 436, 3602
Laporte, C. F. P., Gómez, F. A., Cincotta, P. M., et al. 2015, MNRAS, 453, 2830
Majewski, S. R., Skrutskie, M. F., Weinberg, M. D., & Ostheimer, J. C. 2003, ApJ, 599, 1082
Mateo, M., Mirabal, N., Udalski, A., et al. 1996, ApJL, 458, L13
Myeong, G. C., Evans, N. W., Belokurov, V., Koposov, S. E., & Sanders, J. L. 2017, MNRAS, 469, L78
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 933
Schönrich, R., & Aumer, M. 2017, MNRAS, 472, 3979
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Smith, M. C. 2016, in Tidal Streams in the Local Group and Beyond, Vol. 420, ed. H. J. Newberg & J. L. Carlin (Berlin: Springer), 113
Smith, M. C., Evans, N. W., Belokurov, V., et al. 2009, MNRAS, 399, 1223
Springel, V., Wang, J., Vogelsberger, M., et al. 2008, MNRAS, 391, 1685
Tumlinson, J. 2010, ApJ, 708, 1398
Xue, X.-X., Ma, Z., Rix, H.-W., et al. 2014, ApJ, 784, 170
Yanny, B., Newberg, H. J., Kent, S., et al. 2000, ApJ, 540, 825

Yuan et al.