Star Cluster Simulation

Neelesh C. A.
Dept. of Computer Science Engineering
PES University
Bengaluru, India
neeleshca26@gmail.com

Ninaad R. Rao
Dept. of Computer Science Engineering
PES University
Bengaluru, India
ninaadrrao@gmail.com

***Abstract*—Stars are mostly present as clusters rather than in isolation. Various factors change the course of these clusters and the cluster may dissolve quickly or are present for a longer duration of time. We consider data for a single cluster. Some stars tend to remain in the cluster and some stars leave the cluster. We predict whether a particular star will tend to remain in the cluster or leave the cluster after certain amount of time. It's position/velocity at a particular time are also predicted. The models implemented for classifying a star as whether it left the cluster or not were support vector regression and logistic regression. Due to the imbalance of data, i.e. only 30 stars left the cluster, split was done based on whether a star was an outlier or not. Logistic regression provided a better accuracy and faster execution time. The models for predicting position were a time series model and two regression models. The times series model was ARIMA and the regression models were linear regression and K nearest neighbours regression. Position was predicted and ARIMA gave a better RMSE although execution time was larger. Between linear and K nearest neighbours regression, linear regression had a better RMSE and execution time.**

***Index Terms*—star cluster simulation, predictive analytics, exploratory data analytics, predictive models, time series**

I. Introduction

Depending on a variety of conditions, star clusters may dissolve quickly or be very long lived. The dynamical evolution of star clusters is a topic of very active research in astrophysics. Some popular models of star clusters are the direct N-body simulations [1, 2], where every star is represented by a point particle that interacts gravitationally with every other particle. This kind of simulation is computationally expensive, as it scales as $O(N^2)$ where N is the number of particles in the simulated cluster.

An N body simulation is one where a dynamic system of particles are simulated, interacting with each other through physical forces, such as gravity in our case. The types of particles can range from stars in the universe to atoms in a gas cloud. A star cluster is a group of stars. There are two types of star clusters, globular clusters and open clusters. Globular clusters are large number of stars that are tightly bound by gravity while open clusters are more loose. We are going to work on a globular cluster. Since the lifetime of star clusters can be billions of years, actually observing them is not possible. Hence, there is a need to predict the star clusters at a particular period of time. If it is done for many star clusters, researchers can have better knowledge about the origins and evolution of the galaxy.

At present, star cluster simulations are done on supercomputers that have numerous GPUs and specialized hardware such as GRAPE. Most people do not have access to these resources, and hence the number of people that can work on them are not too many. If we're able to build a suitable model, it would advance the field by a significant amount.

II. Literature Survey

Most of the previous work has been done considering only the aspects of astrophysics and not the analytics point of view.

Aarseth [2] presented an $O(N^2)$ solution where every star is represented as an individual that could interact with every other particle. Here, N is the number of stars in the cluster to be considered. The main principle behind this computation is the Newtons law and any additional potential field if present. The problem is thus a set of non-linear second order differential equations relating the acceleration with the position of all the particles in the system. Once the conditions are specified, model can be constructed accurately. Though this solution could be highly accurate but the computation power is very high and the time it takes to find whether a particular star will remain in the cluster or not could take a really long time.

A binary classifying Deep Neural Network model was built to predict the stability of circumbinary planets in [5]. It is trained on a N-body simulations generated with a data set called REBOUND N-body integrator. This model has an accuracy of around 86 percent, and it does not go below 86 percent even when tightly surrounded by instability.

III. About the Dataset

The data set consists of snapshots of the cluster for 19 different time stamps and each time stamp is separated based on the standard mentioned in N-body units[4]. The data set contains information of the star such as the position along the x, y and z direction and also the velocities in the three directions. It also contains the mass of each star and the id of the star. For the purpose of the simulation, the mass of each star is assumed to be the same across the cluster, i.e. 1/64000. At time t=0(reference time), there are 64000 stars that are present and this number reduces as time passes and at t=19 units, there are 63970 stars. This shows that some of the stars have already left the cluster. We propose to build a model that finds whether a star will remain in the cluster or will leave the

cluster at t=19 and further and we will test on t=19. The result of this experiment will also depend on various other factors such as collision between two stars which would drastically change all the attributes of a given star and whether it will remain in the cluster or not. For the purpose of prediction, we will not be considering such external factors and assume the trend is general. Most of the researches have considered this for the field of astrophysics and we focus this data for the purpose of data analysis and predictive modelling.

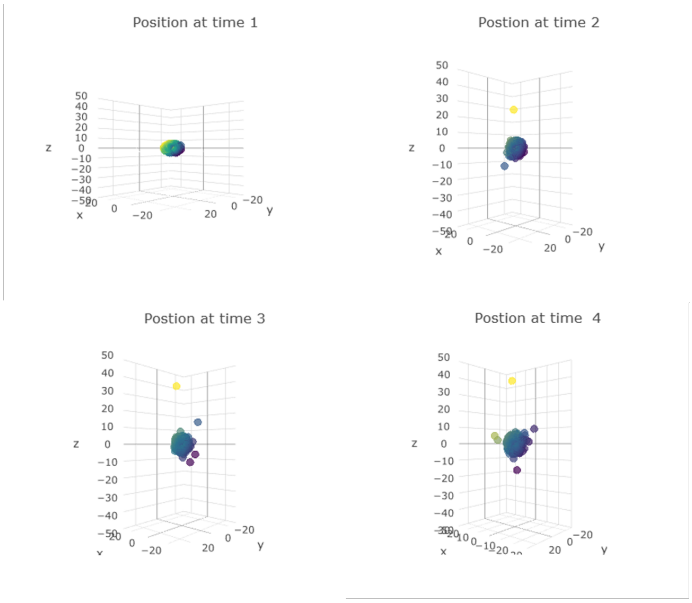

Fig. 1. The change in the position of stars at given timeframes

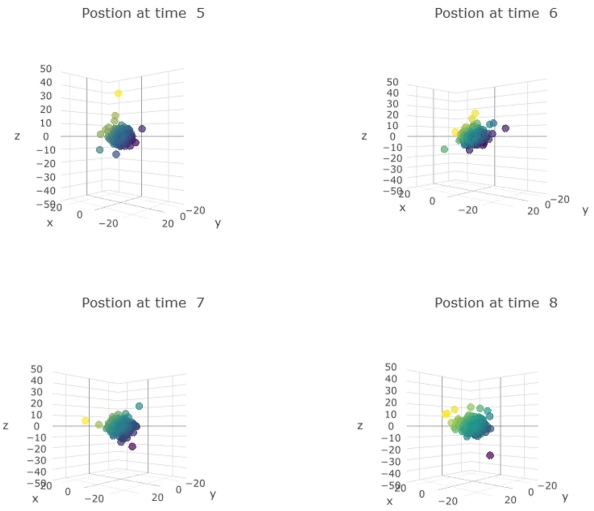

Fig. 2. The change in the position of stars at given timeframes

IV. Problem Statement

There are two main questions to be addressed with respect to star clusters. One is, whether a star will remain in the cluster.

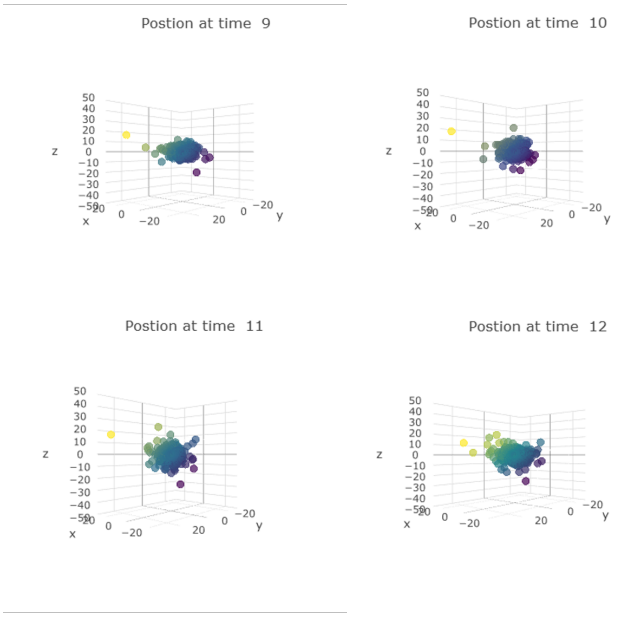

Fig. 3. The change in the position of stars at given timeframes

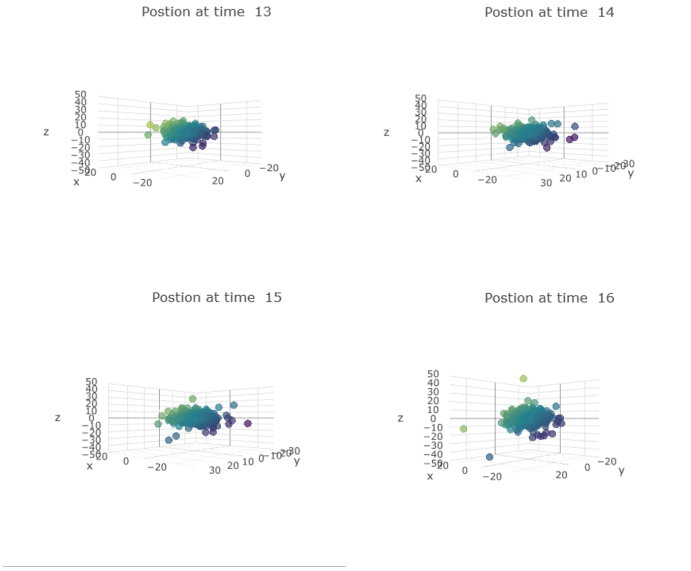

Fig. 4. The change in the position of stars at given timeframes

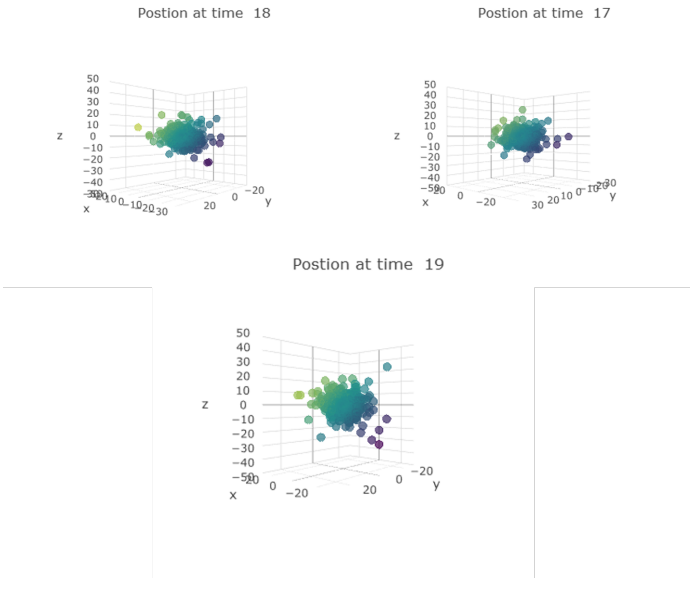

Fig. 5. The change in the position of stars at given timeframes

And if it does leave, when it will do so. The other question is, to predict the position and by extension, the velocities of the stars at a particular point in time. Other details about the cluster like centre of mass of the cluster can be obtained by accurately answering the aforementioned questions. From the figures shown, it is clear that stars tend to change their position every time a collision happens and the figures shows the evolution of the same star cluster for 19 different time-frames. Some of the stars leave the cluster immediately while others might or might not leave i.e. they tend to move within the cluster or reach their escape velocity and leave the cluster.

A faster solution than the direct N-body simulation is the BarnesHut simulation/approximation [6]. The Barnes-Hut simulation considers a group of stars that are far away as one single body, using their centre of mass. This speeds up execution, as only one body has to be looked at, instead of each individual star. The stars that are close by are still looked at as individual entities. The space is divided into an octree, where the root node represents the whole space and its eight children represent the 8 quadrants. Once all the points are inserted, each node contains either 0 or 1 point. The parents contain the centre of mass and total mass of their children. Depending on the parameter that is set for what should be considered long range/short range, the accuracy/computation require varies.

There are other methods of simulation such as the particle mesh method and the fast multipole method. The use of machine learning and deep learning in this field is fairly recent [5]. There is an assumption that direct N-body simulation is the most accurate simulation available. But there can be no way to conclude that due to the observable data being so small. The equations that are used for the simulations may not be able to take into account every possible interaction between the objects, i.e. apart from gravity. A model might be able to recognize features that the formula might not have considered. Hence, the model might be able to create a better simulation, at least with respect to computational power vs accuracy.

The data set was simulated using direct N-body simulation. The time stamps provided for the data set were of a large difference (100 crossing times apart, where one crossing time is of order 10^5).

The problem we seek to address is the computation power required for such simulations. Although they are accurate, the computation power and specialized hardware required for these simulations is not something that is easily available. Other models such as regression, support vector machines etc could be used, but they were not used in this field. The amount of information present in the data set can be considered a constraint as more parameters could be present that influence the movement of stars.

V. Basic Exploratory Data Analysis

Basic exploratory data analysis was done which confirmed our assumptions/gave us information on what models we can use.

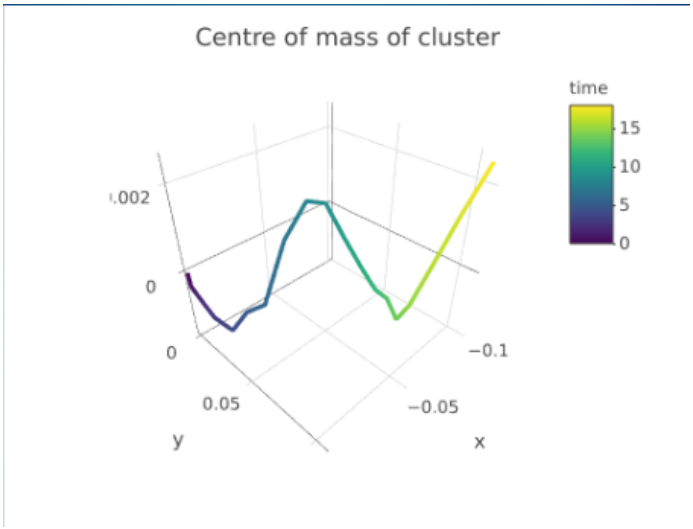

Fig. 6. Movement of centre of mass of the cluster with respect to time

This graph shows no particular trend. This means that the cluster as a whole was not influenced by anything else (larger external cluster for example). The centre of the cluster also moves away meaning that the stars are not in equilibrium.

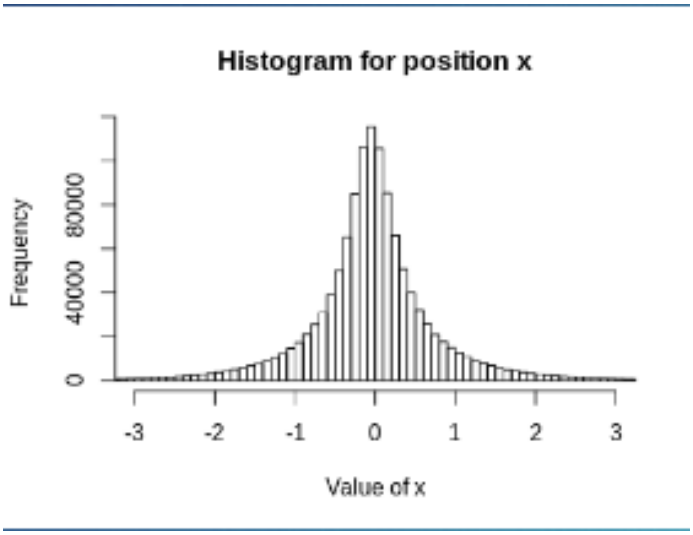

Fig. 7. Histogram of position x

The histogram is normal and does not have any outliers. This means it can be used for regression.

As observed in figure 8, the correlations are all zero. This means no multicollinearity would be present during regression.

VI. Support Vector Regression

Support Vector Regression(SVR) is very similar to the support vector machine(SVM) except that the svr is a regression model. In svm, there are two more lines other than the hyper plane. In SVR, this line is used to predict the outcome of the model.

Fig. 8. Correlation plot

Since there is no label in the dataset that says whether a particular star left the cluster or not, This was marked by considering the outliers as those that do not remain in the cluster.

A binary classifier was made to predict if a star remained in the cluster or left the cluster. The star remaining in the cluster was marked with 0 and that that left the cluster was marked 1. The training data set contained time stamps from 0 to 18 and the test was done on the 19th timestamp. It was a linear model and the prediction was for the final binary value.

The support vector regression model gave a testing accuracy of 95.38% and gave a training accuracy of 95.98%.

The drawback of this model is the speed. To run the svr model for the entire dataset took really long. Since this model was a binary classifier, even if it classified the star correctly, the exact location of the star could not be predicted.

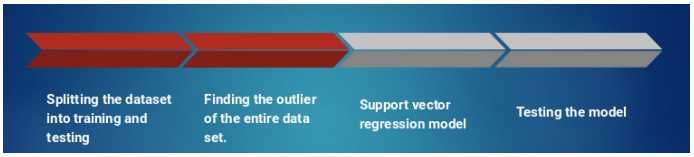

Fig. 9. Block diagram for classifying a star as it having left the cluster or not

VII. Logistic Regression

Logistic regression is a model that is used to predict the value of a variable when it has only two outcomes, i.e. yes or not. Although labels could be added to the stars stating whether they have left the cluster or not, there would be a large imbalance as only 30 stars out of 64000 left the cluster. Hence, outliers were used as the true/false variable.

Since the correlation of the all the variables were 0, all the variables were used in the prediction. Using the model on the train dataset, the accuracy was 99.99 which possibly indicates over fitting. However, a good accuracy of 99.63 was obtained when the model was used on the test dataset. Although this doesn't exclude ovefitting completely, the probability of over fitting being present was low.

The execution time of logistic regression fairly fast when compared to support vector regression.

VIII. Multi Output Regression

Since the ARIMA model could only predict the magnitude of the position, there had to be a model that could give more information. Only the magnitude of the position may not give enough information about the star. Hence models that could predict the individual vectors were needed.

These models were Linear Regression and K Nearest Neighbours Regression. As the correlation of the variables were 0, and a linear relationship is expected between positions and velocity, linear regression seemed suitable. K Nearest Neighbours was used as it is expected that the stars that are close to each other would behave similarly. This is analogous to the barnes-hut simulation which is used in direct N body simulation. Other models such as ridge regressor, random forest regressor and ADA boost regressor were used but they did not give any better results.

Implementing an artificial neural network was also attempted. The artificial neural network was run for epochs 5-30 but the error of the model did not change. This possibly indicates over fitting. Since the training data set was large (64000 stars over 18 time stamps), larger epochs took too long to run.

The train-test split was done as follows. The last time stamp was taken as the test data set and the rest of the time stamps were trained on. Extra attributes like the magnitude of distance and velocity were added to the dataset as well. Since the mass was considered constant for all the stars, it was not for training. The ID of the star was not used either as the ID's were given arbitrary.

There are two important things that could be obtained from this prediction. One being the position and the other being velocity. The direct N-Body simulation is one that is accurate but needs large computing power and specialized hardware to function. But, the regression models could be run on our machines. Prediction for position and velocity were done on two different inputs. One having just the position/velocity vectors and the other having the vectors in addition to their magnitude. But this did not improve results, in fact, the RMSE became larger. This could mean that the model is being overfitted.

Multi linear regression was chosen as one of the models due to it's computation time and it's simplicity. The requirements for regression were satisfied/assumed. There were no outliers as seen in the histogram which means that the line will not be skewed. The correlation of all the variables are 0, meaning no multi-collinearity. The residuals can be assumed to be homoscedastic. The model performed quite well as the computation was fast and the RMSE was fairly low (close to 1).

K Nearest Neighbours was used with a value of 200 for N. The value was obtained by doing an incremental analysis starting with N = 5. After 200, the accuracy did not improve too much but the computation time grew by a lot, which is to be expected. This indicates that the optimal N was 200.

```
The results are the RMSE values
Considering magnitude of velocity in addition to velocity vectors to predict position vectors
Linear Regression :1.1412059155246128
KNeighborsRegressor :1.1440114470573572


Considering only velocity vectors to predict position vectors
Linear Regression :1.1412017234021627
KNeighborsRegressor :1.1436454731971437


Considering only position vectors to predict velocity vectors
Linear Regression :0.40814847129374016
KNeighborsRegressor :0.4090772626901037 5


Considering position vectors and position magnitude to predict velocity vectors
Linear Regression :0.40814895849350197
KNeighborsRegressor :0.4090971448307957
```

Fig. 10. The change in the position of stars at given timeframes

IX. Time Series Analysis

Time series analysis was done on the stars that left the cluster. There were 30 such stars that were not present in the final test data set and these are the stars that left the star cluster at one point of time. Performing the augmented dickey fuller test shows that the data is not stationary. Removing the trend and the seasonality and seeing the graph of the residual, seasonality and trend. Now the acf and the pacf plot were done to find the parameters for the ARIMA model that is being considered. The root mean squared error of the model was 5.69 and if the model was done for the stars that did not leave the cluster the accuracy was 0.79. But this predicts only the magnitude and not the individual distances.

The arima model is highly impractical as there needs to be 64000 models to find the position of all these 64000 stars. This is very slow.

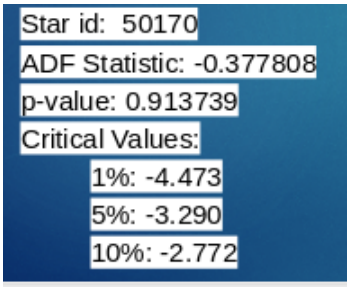

Fig. 11. Augmented Dickey Fuller Test result for a given star

X. Results

As stated previously the experiments were done taking the test-train split and adding features like the magnitude of both the velocity and the distance, results were computed for different models.

The differences are stated as follows. Also, a comparison was done between linear and k nearest neighbours earlier.

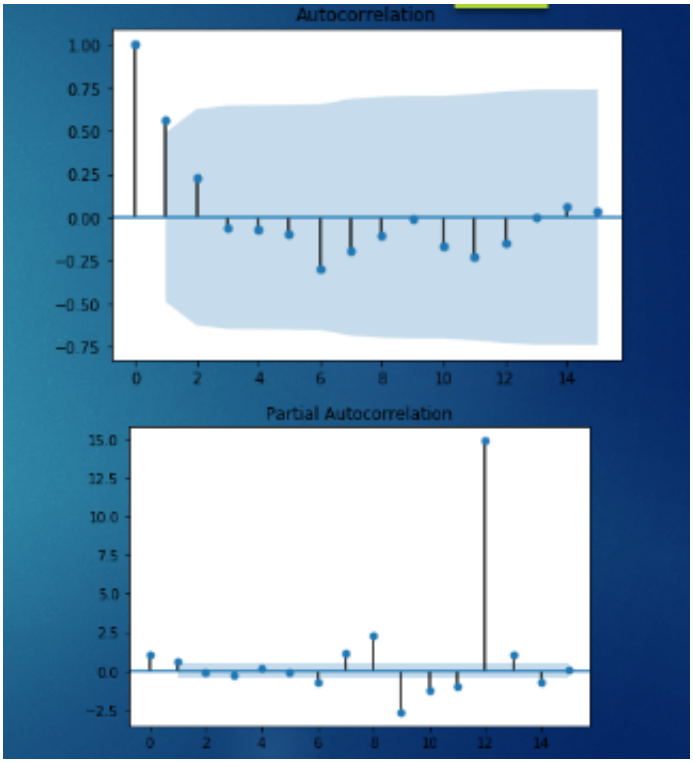

Fig. 12. ACF and the PACF plots of the same star

ARIMA	Multi ouput Regression
Arima was time series analysis.	Based only on present values
Could be done only for magnitude	Could get each vector individually.
Arima was really slow	Relatively fast
RMSE: 0.79	RMSE: 1.41

Fig. 13. The change in the position of stars at given timeframes

Testing Accuracy: 95.37908394559949 Training Accuracy: 95.98391656479217	Training Accuracy: 99.9985239776 Testing Accuracy: 99.6263873691
Slower	Faster
SVR finds the probability	Logistic regression finds log(probability)

Fig. 14. SVR vs Logistic Regression

XI. Contribution

Ninaad- Time series analysis, svr model, visualization

Neelesh - Exploratory data analysis, logistic regression and multi output regression.

References

[1] Heggie, D., Hut, P. 2003, The Gravitational Million-Body Problem: A Multidisciplinary Approach to Star Cluster Dynamics Cambridge University Press, 2003

[2] Aarseth, S. J. 2003, Gravitational N-Body Simulations - Cambridge University Press, 2003

[3] Heggie, D. C., Mathieu, R. D. 1986, Lecture Notes in Physics, Vol. 267, The Use of Supercomputers in Stellar Dynamics, Berlin, Springer
[4] Christopher Lam, David Kipping; A machine learns to predict the stability of circumbinary planets, Monthly Notices of the Royal Astronomical Society, Volume 476, Issue 4, 1 June 2018, Pages 56925697, https://doi.org/10.1093/mnras/sty022
[5] BarnesHut simulation/approximation https://www.nature.com/articles/324446a0