Visualization and Characterization of Users in a Citizen Science Project

Alessandra Marli M. Morais[a] and Jordan Raddick[b] and Rafael D. C. Santos[a]

[a]Brazilian National Institute for Space Research, São José dos Campos, SP, Brazil;
[b]The Johns Hopkins University, Baltimore, Maryland, USA.

ABSTRACT

Recent technological advances allowed the creation and use of internet-based systems where many users can collaborate gathering and sharing information for specific or general purposes: social networks, e-commerce review systems, collaborative knowledge systems, etc. Since most of the data collected in these systems is user-generated, understanding of the motivations and general behavior of users is a very important issue.

Of particular interest are citizen science projects, where users without scientific training are asked for collaboration labeling and classifying information (either automatically by giving away idle computer time or manually by actually seeing data and providing information about it). Understanding behavior of users of those types of data collection systems may help increase the involvement of the users, categorize users accordingly to different parameters, facilitate their collaboration with the systems, design better user interfaces, and allow better planning and deployment of similar projects and systems.

Behavior of those users could be estimated through analysis of their collaboration track: registers of which user did what and when can be easily and unobtrusively collected in several different ways, the simplest being a log of activities.

In this paper we present some results on the visualization and characterization of almost 150.000 users with more than 80.000.000 collaborations with a citizen science project – Galaxy Zoo I, which asked users to classify galaxies' images. Basic visualization techniques are not applicable due to the number of users, so techniques to characterize users' behavior based on feature extraction and clustering are used.

Keywords: Visual Analytics, Visualization, Citizen Science, Collaborative Systems

1. INTRODUCTION

Technological advances allows computer systems to store and exchange a wide amount of data, enabling the creation and use of internet-based systems such as social networks, e-commerce review systems, collaborative knowledge systems and others, where users can collaborate gathering and sharing information for specific or general purposes.

One of the collaborative information sharing approaches are the so-called citizen science networks. In these collaboration approaches, aspects of data collection or analysis are beyond the capacity of the science team, but becomes doable with the involvement of volunteers,[1] namely, members of the public, which are often untrained in the science aspect but can nonetheless collaborate in the effort. The tasks assigned to volunteers are vast and include collaborations like observation of bee pollination, air quality and bird watching. Practiced since at least the 1700s,[2] recent citizen science projects use web resources to attract and involve volunteers, furthering access, collection and the analysis of data, in addition to allowing the exploration of new domains.

Researches in citizen science are concerned to find ways to persuade users to get and stay involved with the project until their object be concluded, decreasing the number of users who leaves the project ("churn rate") and reducing collaboration errors. Other issue of concern is data quality: scientific endeavors require data of very high quality, but citizen science projects intentionally place responsibility for data tasks into the hands of non-experts,[3] although some researchers conclude that data produced by volunteers are as good as data produced

Send correspondence to R. Santos (rafael.santos@inpe.br)

by professional scientists,[2,4] justifying some research on issues and challenges in motivation of citizen scientists[5] and proposing mechanisms to enhance the quality and trust of citizen science data.[6–8]

Understanding the behavior of volunteers may help to overcome some issues and challenges of citizen science projects, helping understand reasons for users' abandonment of the project (and possibly suggesting strategies to prevent this), categorize users accordingly to different parameters with the aim to increase data quality by prioritizing or penalizing some collaborations, facilitate their collaboration with the systems, design better user interfaces, and allow better planning and deployment of similar projects and systems.

In this paper we present some visualization techniques that can be used to shed light on some aspects of the behavior of volunteers in a citizen science project. Characterization of the users' behavior is important so we can better motivate users to join the project and avoid issues that can cause users to leave it.

This paper is divided as follows: Section 2 presents the basic concepts of citizen science and the users' data that these projects generate. Section 3 presents some visualization concepts and techniques appliable to this domain. Section 4 shows a simple technique that can highlight points in time that experienced large numbers of users joining and/or leaving the project, making possible the mapping of those points in time to real-life events. Section 5 shows a simple neural network-based technique to visually segment users' groups and make possible the labeling of those groups. Finally, section 6 presents the conclusions and directions for future work.

2. CITIZEN SCIENCE AND THE GALAXY ZOO PROJECT

The term "citizen scientists" refers to volunteers who participate as assistants in scientific studies.[9] The assistance is usually performed by creating data from observations. For example, volunteers may monitor wild animals and plants, collect in the study of air quality and help in the production of galaxies catalogs, as happens in Galaxy Zoo project. Such volunteers are not paid for their assistance, nor are they necessarily even scientists. For example, for ecological research citizen science projects or projects linked to nature the "citizen scientists" most are amateurs who volunteer to assist ecological research, because they love the outdoors or are concerned about environmental trends and problems, and want to do something about it.[9]

Data-based scientific researches require very high quality data and the fact that most citizen science projects intentionally place responsibility for creating data into the hands of non-experts[3] may seem antagonistic. This brings a prejudiced view about the scientific validity of these projects by some scientists.[2] Thereby, a common doubt on using citizen scientists is why depend on amateurs, who may make mistakes, not fully understand the context of the study and consequently may provide unreliable data? One reason is economic: it is possible to conduct data collection surveys without the financial resources required for the project, particularly without the need to hire many assistants researchers to do what the volunteers do.[9]

Scientific data is currently doubling every year,[1,10] while the number of professional scientists available to interpret the data grows much more slowly. Therefore, scientific projects in which some aspect of the data collection or analysis is beyond the capacity of the science team, but that becomes doable with the involvement of volunteers, should have their scientists prepared to scrutinize the data carefully and willing to discard suspect or unreliable data.[9]

Furthermore, it is common that scientists strive to engage the volunteers so they can learn more about the research in which they are collaborating.[11] This approach has demonstrated that efforts of citizen science projects often produce data with scientific validity and high quality.

The Galaxy Zoo project is a good example of engagement success among scientists and volunteers to produce data with quality. For most of the twentieth century, morphological catalogues of galaxies were compiled by individuals or small teams of astronomers, but modern surveys, like the Sloan Digital Sky Survey (SDSS),[12] containing data from millions of galaxies make this approach impractical.[13]

Started in June 2007, the Galaxy Zoo project asked users to look at pictures of galaxies and report on their morphological features, allowing the classification of nearly one million galaxies.

A website was designed to recruit users and collect the data provided by them. Visitors to the site were asked to register and read a brief tutorial, and after that were submitted to a simple multiple-choice test with selected galaxies from the SDSS which were previously classified by team members. Volunteers with a low degree

of agreement were rejected, while those who correctly classified 11 or more of the 15 galaxies on the test were allowed to continue and become a volunteer.[4]

Some of the methods used by the research team to keep volunteers engaged were the use of social networks, maintenance of a blog, monitoring and answering users' questions on the projects' forum, reporting on the findings to the users as well as explicit acknowledgments by their collaboration in scientific papers, published by the team. Additionally some efforts to understand the volunteers' motivations and characterize them were done,[5] with interesting results.

The volunteers' generated data was not used as-is: some filtering was performed, and data quality parameters were estimated in order to develop a galaxy morphology catalog.[4,13] To demonstrate the catalog's quality, studies were conducted by comparing them with catalogs compiled by teams of astronomers. The results were considered satisfactory.[4]

Several scientific articles and studies were created as a direct consequence of the committed effort by the team (researchers and volunteers). The list of publications and results can be found in the Zooniverse website (`https://www.zooniverse.org/`), a portal that hosts several citizen science projects.

3. VISUALIZATION OF USERS' BEHAVIOR

Visualization usually described as a mapping of data to a visual representation which amis to improve the interaction of users and data. Thanks to the human visual perception ability, a visual representation of a dataset may present more information than the data in its raw form, be it numerical or textual.

Visual data representation allows understanding complex systems, making decisions, and finding information that otherwise might remain hidden in the data.[14] However, there is no universal technique that enables interpretation and interaction as desired on any dataset – different visualization techniques can be used for visualizing different data types. Some visualization techniques are specially designed to support one specific data type, others are more general, allowing the visualization of a range of data types.[14]

Selection of a visualization technique depends largely on the task being supported and it is still a largely intuitive and *ad hoc* process. Taxonomies and classification schemas only provides some initial insight on which techniques are oriented to certain data types.[15]

When choosing a good visual representation for the data in a specific task, one must chose a technique or representation that presents good recall and easy understanding, avoiding complex captions and undesirable visualization characteristics like occlusions and line crossings, that might appear as an artifact limiting the usefulness of the visualization technique.

The nature of the data used in this study precludes the use of some basic visualization techniques, so specific techniques and data representation schemas were used to allow its visualization. In section 3.1 we describe the data and its nature. Sections 4 and 5 presents the techniques we used and the results of the visual interpretation of the data.

3.1 Data used for visualization

Volunteers on the Galaxy Zoo project were presented with images showing galaxies, and had to answer simple questions about the galaxies' shapes and orientation (clockwise or counterclockwise). After clicking on the option they believe is the most appropriate, a new galaxy image is automatically displayed and once again the system is ready to record a new collaboration.

For each classification, the system stores the volunteer identification, the galaxy identification, the timestamp and the classification chosen by the volunteer. We want to study similarity in volunteers' behavior, in other words, search for patterns in the way volunteers collaborate with the project. It is not our goal infer about the quality of classification or attempt to identify volunteers directly or indirectly, that's why the user IDs were anonimyzed. For this study we used only the collaborating data itself: a small dataset with the user ID, the total number of classifications done, the first and last day he or she used the system and the daily number of classifications.

The data was obtained from records from the Galaxy Zoo I project, and cover the period from its launch in July 8, 2007 until July 7, 2012 (1,822 days). During this interval 146,669 volunteers collaborated with the

project, averaging 540.64 collaborations per user. A single user contributed with 1,220,067 classifications in approximately eighteen months, and two other users had more than half a million collaborations.

During the 1,822 days considered in this study, the highest daily amount of collaborations happened on July 13, 2007, with 1,398,722 collaborations, while the daily average was approximately 43,521 collaborations.

Other basic information about the dataset is shown in the next tables. Table 1 summarizes the distribution of volunteers per number of collaborations, divided into 7 arbitrary intervals. Table 2 shows the distribution of the collaboration span of the volunteers in total years of activity in the project.

Table 1: Collaboration Interval

Interval	**0-10**	**10-100**	**100-1,000**	**1,000-10,000**	**10,000-100,000**	**100,000-1,000,000**	**>1,000,000**
Percentage	15.88%	45.35%	29.74%	8.32%	0.69%	0.03%	0.00%
Volunteers	23,291	66,515	43,619	12,196	1,009	38	1
Accumulated	23,291	89,806	133,425	145,621	146,630	146,668	146,669

Table 2: Collaboration Period

Interval	**0-1 years**	**1-2 years**	**2-3 years**	**3-4 years**	**4-5 years**
Percentage	96.71%	3.01%	0.14%	0.10%	0.04%
Volunteers	141,847	4,409	205	154	54
Accumulated	141,847	146,256	146,461	146,615	146,669

4. THE BIG PICTURE – VISUALIZING BEHAVIOR OF 140,000 USERS

Many techniques for data visualization has been developed, but not all of those are adequate to the flood of data that can easily be obtained nowadays.[16] In most of the proposed approaches, the number of data items that can be visualized on the screen at the same time is limited (in the range of 100 to 1,000 data items) and increase this limit has been a declared research goal.[17] Increasing the amount of data that can be used in a visualization task is also an interesting research topic due to the impact of the data increasing in the visualization tasks' speed, complexity, hardware and software requirements and viability itself.[18]

We would like to visualize some time-related events that affect or involve many users at a time. For this we must consider all the volunteers during all the time of their collaboration, giving a "big picture" view of the interaction of the users with the data collection system.

Pixel-based or icon-based visualization techniques uses a small visual region (for pixels) or geometrical representation (for icons) to represent each individual data on a dataset. In our case we have 146,669 users that interacted with the data collecting system during 1,822 days. Depending on the patterns we want to visually identify in this data, we could end up with a large, unwieldy visual representation. For example, a simple X-Y or 3D histogram plot to show the behavior of each user in time could be represented by a plot with $146,669 \times 1,822$ graphical elements, which while feasible would not be easy to interact with, and worse, would present the data a slice at a time (e.g. using pages or interactive scrolling techniques), making it hard to see the "big picture". We could scale down this kind of data plot but it could cause occlusions, that could hide some significant information from the viewer.

In order to display the whole dataset at once and keep the relevant information visible, we devised a specific icon-based visualization schema. In this schema circles are used as visual icons and are distributed in a XY-plot, where values on the Y axis correspond to the first day of collaboration and values on the X axis the last day of collaboration. Each circle in a XY coordinate represents a group of volunteers that joined and left the project in a specific combination of two days, with the radius corresponding to the number of volunteers in that particular group and the color of the circle corresponding to the sum of classifications performed by that group.

The XY-based representation allows us to verify the volume of users that entered or left the project in a date or ranges, while the circles' attributes allows the visualization of volume of users and collaborations in a specific day or range of days. This visualization schema is similar to a three-dimensional histogram or three-dimensional XY plot, but with the added advantage of drawing attention to large shapes, perceptually groups of shapes and streaks (linearly-oriented groups of shapes).

With this visualization schema we plotted the data corresponding to the 146,669 volunteers which done 79,295,697 classifications in 1,822 days. The results are shown in Figure 1*. Labels for the icons shown in Figure 1 are shown in Figure 2.

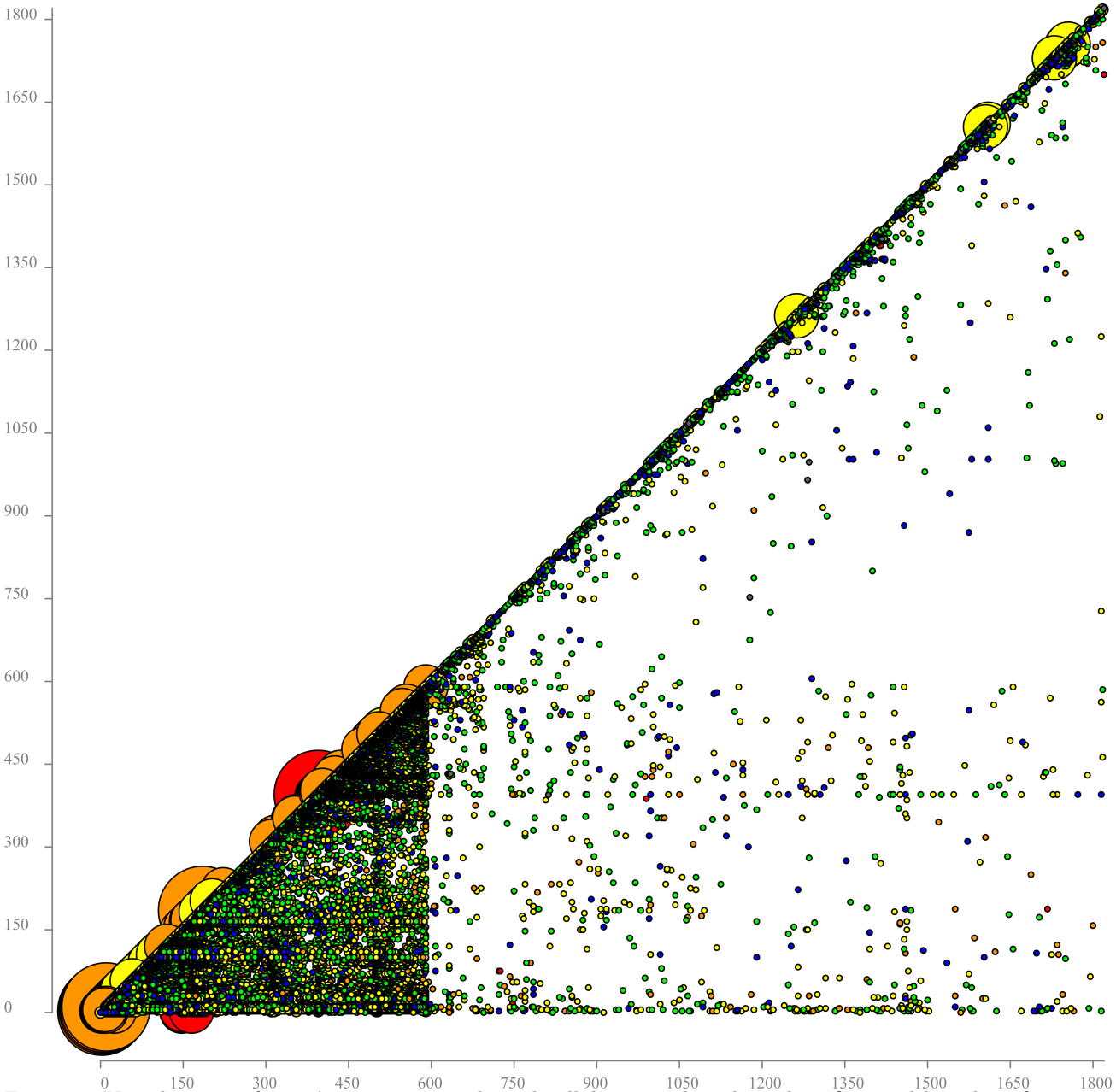


Figure 1: Visualization of users' groups size and total collaboration based on their first and last day of activity in the project.

From the visualization shown in Figure 1 it is possible to observe some interesting features about the data and then investigate the reasons for their occurrences. The main visible feature is the division of the points in two regions with distinct density, separated around the 600 days' mark. Other visible features are the formation of two horizontal lines close to 0 and 400 days, showing that volunteers who entered near these dates abandoned

*Color versions of these figures can be obtained through the contact author.

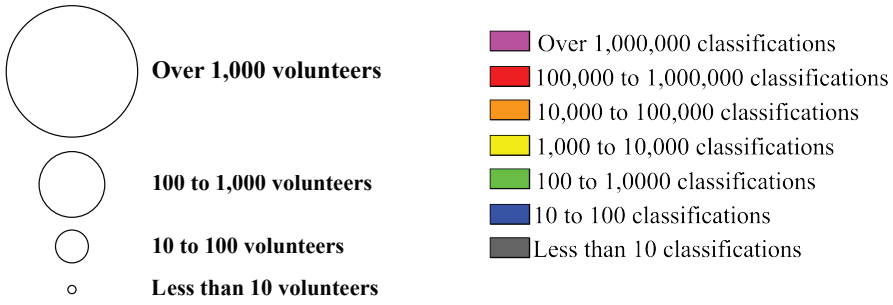


Figure 2: Labels for Figure 1.

the project in a gradual and persistent way, throughout the 1,822 project days. Finally, it is possible observe that in the region corresponding to volunteers with short period of collaboration (the main diagonal on the plot), there are circles with radii larger than most of the others in the plot. In an attempt to assign meanings to these occurrences, we analyzed entries on the projects' blog, forum and newsletters, maintained by members of the project, for insights in what may have caused these visual patterns.

The vertical division line around day 600 on Figure 1 indicates that several users stopped collaborating with the project on or around that day. On February 18, 2008 (day 591 of the project) the project's team members informed their volunteers about the launch of a new version of the project, the Galaxy Zoo II. Thus it is believed that the formation of two regions may be justified by the possible migration of volunteers from Galaxy Zoo I to Galaxy Zoo II.

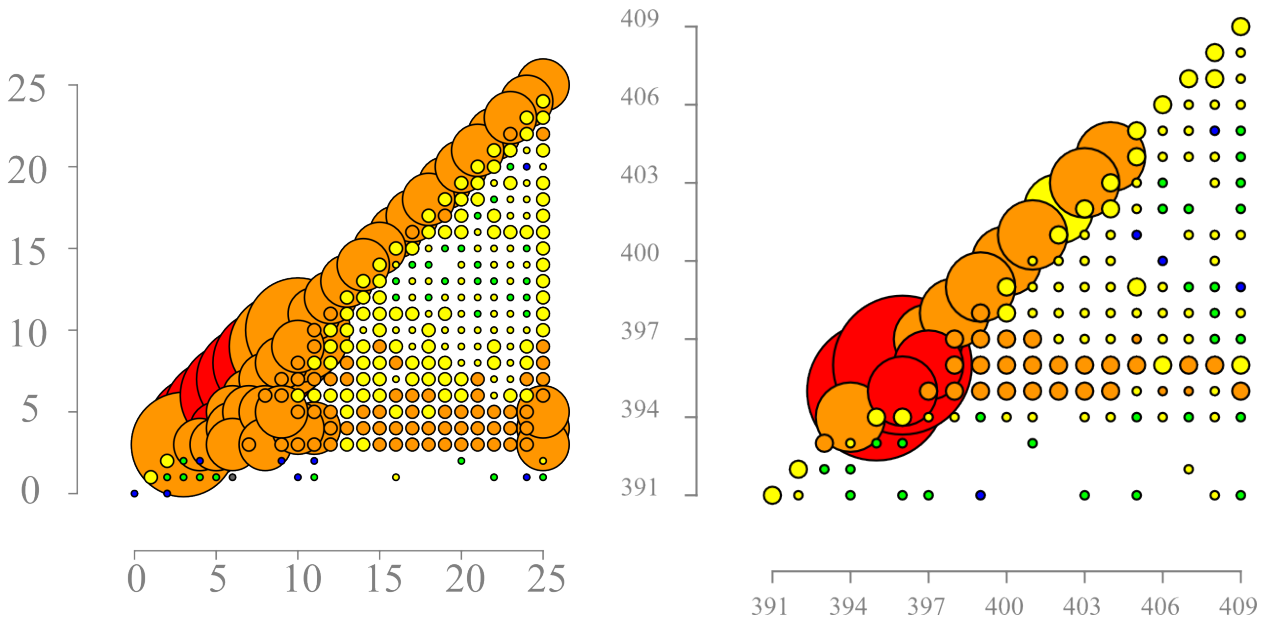
Horizontal streaks in the plot in Figure 1 means that a significant amount of users joined the project on the day corresponding to the height of the line and left the project slowly and on different days, i.e. they kept collaborating for some time. The reason for the joining by several users on the same day (or short days' interval) may be explained by sudden interest by several users, which by its turn may be caused by a number of factors. For example, on July 11, 2007, (third day of the project), Chris Lintott (an astrophysicist working in the Department of Physics at the University of Oxford and one of the main proponents of the Galaxy Zoo project) announced the project in a BBC radio program. The news quickly spread through the media, dramatically increasing the volunteers participation.[5] Figure 3a zooms in the first 25 days of the project, showing that for the first three days of the project there were very few volunteers joining it, while the number exploded for several days after the announcement.

Another event that may have increased the project's visibility occurred near day 395, due to the discovery of an astronomical object of unknown nature dubbed "Hanny's Voorwerp", which discovery was reported in several news outlets, including CNN, resulting in a new wave of volunteers joining the project. Figure 3b shows the region of the plot corresponding to day 395, evidencing a large number of volunteers that joined the project in that day or shortly afterwards.

Other details related users joining and leaving the project can be better visualized if we zoom in on Figure 1 showing only the first 625 days of the project, as shown in Figure 4. In that Figure it is possible to observe new features in the data, for example, more well-defined horizontal and vertical lines, corresponding, respectively, to a relatively large amount of users entering and leaving the project in a particular date. Some of the horizontal lines also have large circles associated, positioned on the main diagonal, meaning that not only several users joined the project in those days (leaving shortly) but they also give a relatively large number of collaborations.

It wasn't easy to find explanations for some of the horizontal and vertical lines for the first 625 days of the project, but some searches on the forum and blog of the project revealed some dissatisfaction of the users with a temporarily change on the data shown to the users, around the beginning of December 2007 (day 146 of the project) – at that time the Galaxy Zoo's team conducted a scientific study about the influence of color on the galaxies' classification by the volunteers.

Simple analysis of the volunteers' collaborations reveals that 91.78% of the volunteers were active on the data collection site between days 0 and 600, contributing with 90.52% of the total classifications.



(a) First 25 days of the project.

(b) Hanny's Voorwerp effect.

Figure 3: Details of horizontal streaks in Figure 1.

Of these, 62.82% of the users can be seen as curious who visited, cooperated for some short time and left the project on the same day, making a total of 6,256,407 classifications, 7.8% of the total project ratings.

The 37.18% remaining users made a total of 65,525,422 classifications, corresponding to 82.63% of total. Volunteers who entered in this period and remained working after the launch of Galaxy Zoo II, a total of 839 volunteers, made 4.94% of the total classifications. Finally, 11,218 volunteers, 7.64% of the total, began their collaboration after day 600 and totalized 3,593,149 classifications, 4.53% of total.

5. VISUALIZATION WITH THE KOHONEN SELF-ORGANIZING MAP

Another way to extract information about the users in a citizen science project is by separating the users in groups that exhibit approximately the same behavior – in other words, by grouping users with similar features so we can consider their behavior to be more or less the same. Grouping is often done by clustering algorithms, a technique used in data mining applications with several different algorithms and implementations.

We want not only to cluster the users in similar groups, but also be able to view these groups in order to infer behavior from the groups. A well-known algorithm that can be used for this is the Kohonen's Self-Organizing Map,[19] also known as SOM. The basic SOM is composed of a lattice of vectors or neurons that will, after training, represent the original input data vectors distributed in the lattice in such a way that the topology of the input data is preserved in the lattice – in other words, input data vectors that were similar (or close) in the original feature space will be close (i.e. represented in the same or in a nearby neuron) in the SOM's lattice. Since the lattice has few dimensions (traditionally two) it can be used to group together vectors that are similar in regard to the input feature space.

The basic SOM uses as input the size and type of the lattice, the input data vectors and some traning parameters and convergence verification strategies. Each neuron on the lattice has the same numerical dimensions as the data that will be presented to the algorithm, and before starting the neural network training these neurons are populated with random values. The training steps present one data vector at a time to the network, identifying which neuron is closest, in feature space, to the data vector present (this is called the "winner neuron"). The neuron (and often its neighbor neurons) is updated in order to get closer (again in feature space) to the data,

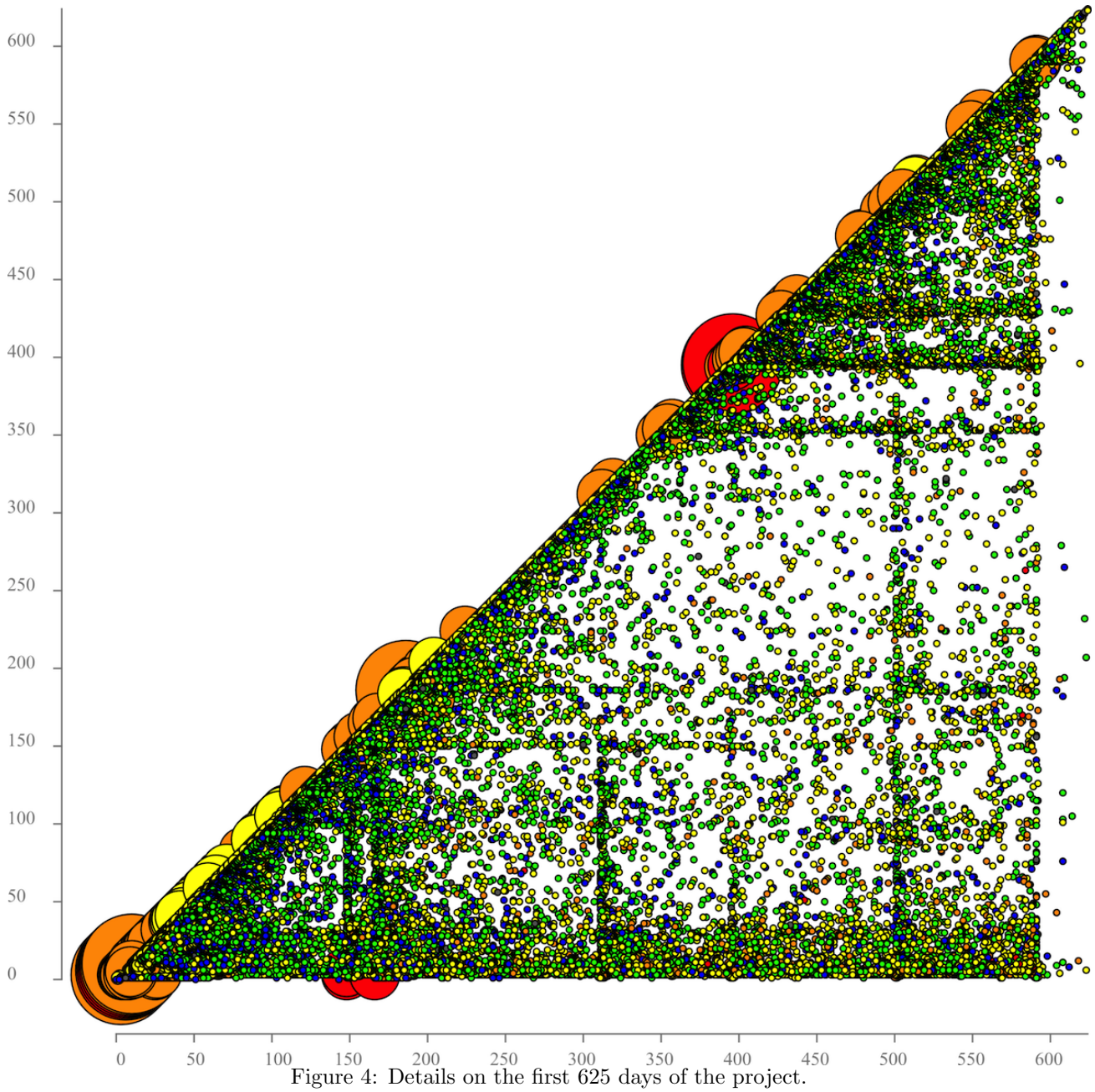


Figure 4: Details on the first 625 days of the project.

and another data vector is presented to the network. These steps are repeated while slowing decreasing the size of the neighborhood around the winning neuron and the amount of change that is done to the affected neurons.

Several variations on the network structure and training strategies are possible[19] but for our purposes the interesting feature of the SOM is the final lattice of neurons that were trained to represent groupings of the data vectors. Since by the algorithm's definition data that is similar in the input feature space is organized in the lattice, neighborhood cells in the lattice represent patterns that are close in the original feature space. If we use the neurons' data in the lattice to present some visual information, we can perceive groupings visually and identify patterns and outliers. The SOM can be used, then, as a specialized icon-based visualization technique, with icons' appearance determined by the values on the trained neurons.

The Self-Organizing Map can be used with any numerical data vector, therefore we could use it to group

users based on their daily activity with the Galaxy Zoo project. Doing so would create clusters of data with users with similar activity patterns grouped together, but visualization would not be efficient: each icon would be very large and visual interpretation could be unwieldy when a lattice of those large icons is presented to the user. Additionally, the large number of dimensions could also cause unwanted effects due to the so-called "curse of dimensionality" that affects clustering techniques based on distance, such as the SOM.[20]

In order to apply the Self-Organizing Map to the data pertinent to this study, we've devised a feature vector with seven attributes that will be calculated from the original data, but considering only users that had 600 or more days in which they could collaborate and interact with the system. In other words, we discarded data from users which joined the project late (in order to be able to calculate some of their features), but preserved data from the users who joined it early on. Doing so we expect to avoid imbalances on the features' values. By eliminate the users which had less than 600 days of recorded interactions we gave up on using 4.68% of the total users and 3.67% of the total collaborations.

The first attribute measures how much a volunteer dedicated him/herself to the project considering the 600 days' window, and is calculated as the range of days (i.e. last minus first) we had registered activity for that volunteer divided by the total period under consideration (600) – values close to zero indicates the user joined and apparently left shortly thereafter, while values close to one indicates collaboration during all or almost all the 600 days.

In order to evaluate the number and frequency of collaboration from that user during that period we devised several other attributes: one that measures the assiduously of the user by getting the ratio of days when we had registered activity divided by the range of days (last minus first) – low values indicate sporadic participation. The next attribute complements the two first attributes, being the division of days of effective collaboration by 600. Finally, another attribute corresponds to the ratio of maximum collaborations in a day by the total of collaborations, which will present values close to one for volunteers who did all the collaborations in a single day, being also an indirect measure of dispersion.

Other three attribute were calculated: the number of collaborations of an user over his/her average and the absolute range of days the user collaborated with the project. Some of those attributes are redundant or may present values inversely proportional to others, but this was done in order to elicit visual patterns on the icons used on the lattice. The last feature is simply the logarithm (base 10) of the total collaborations of the users.

The data was processed in order to filter the eligible users and their feature vectors were calculated to be used in the SOM's training. The network was training using a 15×15 lattice, with adequate training parameters that ensured that all vectors were passed through the network at least 150 times. The visual results of the trained network are shown in Figure 5.

Figure 5 shows 15×15 patterns corresponding to users groups' prototypes as created by the SOM neural network. Each prototype is symbolized by a short plot with several vertical axes and a polyline which crosses those – each vertical axis correspond to one of the attributes, and the position the polyline crosses the axis corresponds to the value for that axis and prototype. This design was inspired by the popular Parallel Coordinates visualization technique.[21] Over each icon we print the number of data points (i.e. users) which can be considered closer (in feature space) to that prototype.

The SOM does not work as a clustering technique: the prototypes that are generated may or not have data points associated to them, and several visually and numerically similar prototypes may be present in the lattice, as we can see in Figure 5. What makes the SOM an interesting visualization tool for multidimensional data is the possibility to identify some regions with similar prototypes which corresponds, usually, to large amounts of data that are similar to the prototypes, without the need to define a specific number of clusters. Using this concept we identify in the lattice shown in Figure 5 several well-represented prototypes with easily mappings to our dataset and expected users' features:

- The rightmost side of the lattice (and to a smaller extent the bottom-right part of it) shows similar patterns which we dubbed "the curious": users who joined the project and did most of the collaboration in one or few days, abandoning the project shortly afterwards.

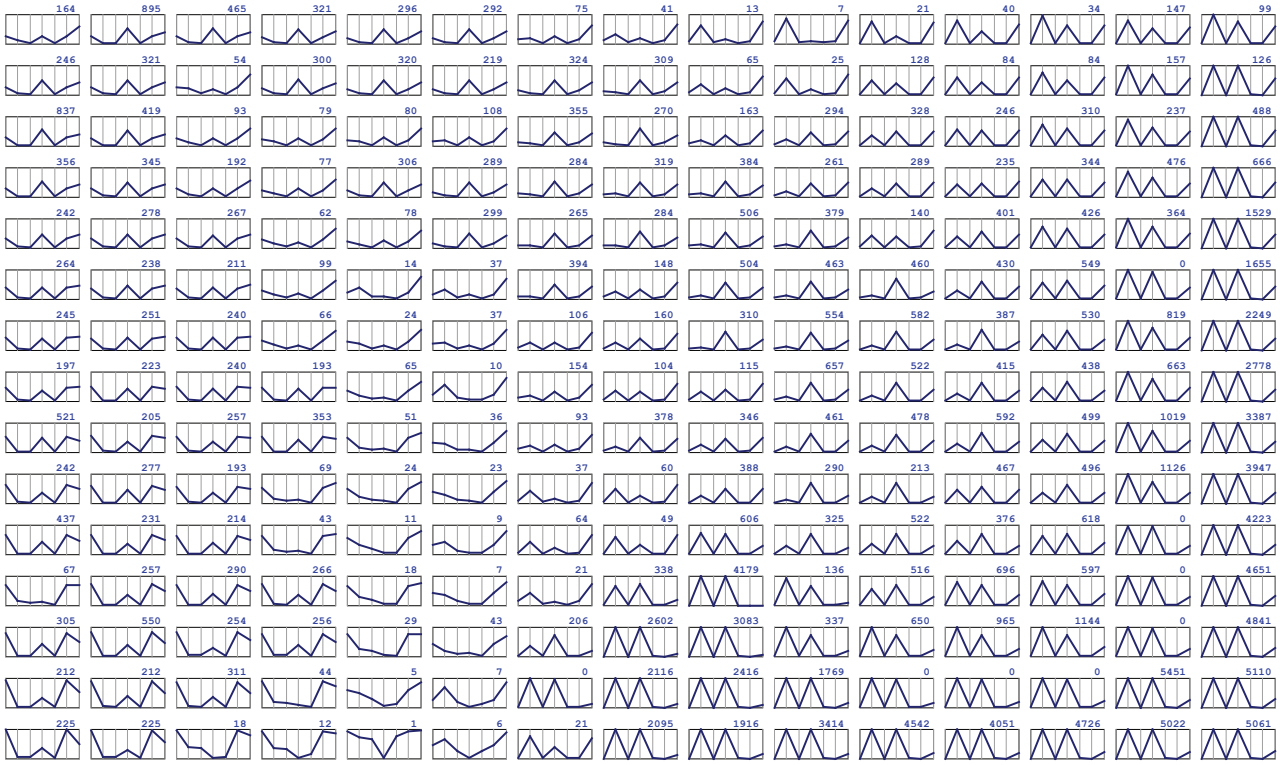


Figure 5: Trained SOM network for prototype's visualization.

- The leftmost side of the lattice contains patterns for users we dubbed "curious with potential to become a regular", which had relatively long collaboration intervals with long intervals between collaborations. Our interpretation is that those users were attracted by the project, collaborated with it for a while, left and then returned again after some time. Frequent collaborators but with long intervals between collaborations also fit this pattern.
- Some patterns on the top right part of the lattice (close to the "curious" ones) were dubbed "the dedicated", users who were assiduous with collaborations distributed along the interval of collaboration.

There are other possible groups of users that can be labeled by visual investigation of the lattice shown in Figure 5, but were outside of the scope of this work. Potentially more interesting is the fact that are patterns with few data points associated to those who may also be significant or representative of a different type of user profile, which can be investigated in a future work.

6. CONCLUSIONS AND FUTURE WORK

In this paper we explored some techniques to visualize different aspects of the behavior of users in a citizen science project. Using simple techniques it is possible to identify patterns of user interaction with the website used to collect data for the project and map those to real-world events that may have impacted on the interest of the users by the project.

Further analysis may lead to more insights on what cause users to get motivated (or not) to collaborate with a citizen science project, so strategies for user motivation and retention may be devised. Other visualization techniques can be used to segment users in different profiles, so different strategies could be used depending on the user's profile.

Future work on this line will be dedicated to other ways to view all data at once in order to illustrate specific aspects of the users' interaction with the data collection site. Other ways to represent icons in a SOM lattice for users profiles' visualization will als be researched.

This research used fully anonymized data from the Galaxy Zoo users – only the random ID and timestamps were used. We expect in the future to be able to use more data (e.g. correct votes, users' self-described data) while keeping users anonymous in order to better characterize them.

ACKNOWLEDGMENTS

Alessandra M. M. Morais would likes to thank the Brazilian National Research Council (CNPq) for financial support for the "Evaluation and Implementation of Medium-Sized Scientific Database Management Systems" project (process number 300227-2012/4).

REFERENCES

[1] Raddick, M. J., Bracey, G., Carney, K., Gyuk, G., Borne, K., Wallin, J., Jacoby, S., and Planetarium, A., "Citizen science: status and research directions for the coming decade," *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey* , 46P (2009).

[2] Droege, S., "Just because you paid them doesn't mean their data are better," in [*Proceedings, Citizen Science Toolkit Conference. Cornell Laboratory of Ornithology. www. birds. cornell. edu/citscitoolkit/conference/proceeding-pdfs*], (2007).

[3] Prestopnik, N. R. and Crowston, K., "Gaming for (citizen) science: Exploring motivation and data quality in the context of crowdsourced science through the design and evaluation of a social-computational system," in [*e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*], 28–33, IEEE (2011).

[4] Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al., "Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey," *Monthly Notices of the Royal Astronomical Society* **389**(3), 1179–1189 (2008).

[5] Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., and Vandenberg, J., "Galaxy zoo: Exploring the motivations of citizen science volunteers," *Astronomy Education Review* **9**, 010103 (2010).

[6] Soares, M. D., *Employing citizen science to label polygons of segmented images*, PhD thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos (2011-06-06 2011).

[7] Alabri, A. and Hunter, J., "Enhancing the quality and trust of citizen science data," in [*e-Science (e-Science), 2010 IEEE Sixth International Conference on*], 81–88, IEEE (2010).

[8] Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K., "Mechanisms for data quality and validation in citizen science," in [*e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on*], 14–19, IEEE (2011).

[9] Cohn, J. P., "Citizen science: Can volunteers do real research?," *BioScience* **58**(3), 192–197 (2008).

[10] Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A., DeWitt, D. J., and Heber, G., "Scientific data management in the coming decade," *ACM SIGMOD Record* **34**(4), 34–41 (2005).

[11] Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., and Shirk, J., "Citizen science: a developing tool for expanding science knowledge and scientific literacy," *BioScience* **59**(11), 977–984 (2009).

[12] Szalay, A. S., "The sloan digital sky survey," *Computing in Science and Engineering.* **1.2**, 54–62 (1999).

[13] Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., et al., "Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies," *Monthly Notices of the Royal Astronomical Society* **410**(1), 166–178 (2011).

[14] Keim, D. A., "Visual exploration of large data sets," *Communications of the ACM* **44**(8), 38–44 (2001).

[15] Ferreira de Oliveira, M. C. and Levkowitz, H., "From visual data exploration to visual data mining: A survey," *Visualization and Computer Graphics, IEEE Transactions on* **9**(3), 378–394 (2003).

[16] Keim, D. A. and Kriegel, H.-P., "Issues in visualizing large databases," in [*Proc. Conf. on Visual Database Systems (VDB'95), Lausanne, Schweiz*], 203–214 (1995).

[17] Treinish, L. A., Butler, D. M., Senay, H., Grinstein, G. G., and Bryson, S. T., "Grand challenge problems in visualization software," in [*Proceedings of the 3rd conference on Visualization'92*], 366–371, IEEE Computer Society Press (1992).

[18] Unwin, A., Hofmann, H., and Theus, M., [*Graphics of large datasets : visualizing a million / Antony Unwin, Martin Theus, Heike Hofmann. [electronic resource]*], New York : Springer (2006). Includes bibliographical references (p. [251]-261) and indexes.

[19] Kohonen, T., [*Self-Organizing Maps*], Springer, 3rd ed. (2011).

[20] Radovanović, M., Nanopoulos, A., and Ivanović, M., "Hubs in space: Popular nearest neighbors in high-dimensional data," *J. Mach. Learn. Res.* **11**, 2487–2531 (Dec. 2010).

[21] Inselberg, A., [*Parallel Coordinates – Visual Multidimensional Geometry and Its Applications*], Springer (2009).