Resolving Conflicts in Anthropogenic Space Object Data Through Weight Distribution Networks with Embedded Data Curation

Nevan F. Simone
The University of Texas at Austin
Dr. Maria Esteva
The University of Texas at Austin
Dr. Moriba K. Jah
The University of Texas at Austin

1. ABSTRACT

We introduce a novel system to resolve conflicts and overlaps in data about Anthropogenic Space Objects (ASO's) in ASTRIAGraph. ASTRIAGraph is a knowledge system to explore Space Domain Awareness (SDA). It contains data about ASOs provided by different collections. While data may be generated with the best knowledge and instrumentation available to each collection, due to different modeling methods, ways of labeling data, or observational and human errors, each collection may provide overlapping or conflicting data about the same ASO. To achieve transparency, predictability, and accountability in SDA, conflicts in the data must be resolved. To address this problem, we devised a data-curation-informed Weight Distribution Network (WDN) method to suggest which collection is more reliable in providing a given field value for an ASO. WDN adapts the PageRank algorithm (PR) to assign and distribute weight across the data fields present in the different collections. Rooted in data curation best practices involving: data completeness, coincidence between field values, and consistency of values over time as metrics of collections' reliability, the final weights provide a basis to resolve conflicting values. We demonstrate the method's capability, and the work that lies ahead to improve it with the goal of enabling more precise ASO identification and characterization.

2. ASO RESOLUTION AND WEIGHT DISTRIBUTION NETWORKS (WDN)

Earth's orbital environment contains a growing field of objects that range in sizes from pins to school buses. Run by different government agencies, ONGs, and private companies, sites that track these objects are spread across the globe. Tracking technology is only improving, and as more sites are capable of tracking smaller objects and higher orbital regimes, the number of cataloged objects increases. Results depend on the source of observations and the algorithms used to compute orbit estimates from those measurements, so many distinct orbital states can be obtained for individual ASOs. While redundant, data about ASOs is not always complete. Characteristics such as physical traits, owner/operator, purpose, and others that are mostly invariant over time are not known across all collections. Since methods currently establish ASO uniqueness based on consistent observations and orbit predictions, most collections are maintained on tracking or on orbit data. When multiple collections are brought together in a knowledge system like ASTRIAGraph [1], a knowledge system to explore Space Domain Awareness (SDA), the factors mentioned above lead to gaps in fields from each collection, and differences between values of any field. These differences lead to ambiguities for those studying the orbital environment. In turn, the scope and size of the collections makes resolving these differences difficult. For example in ASTRIAGraph some of the collections are updated on a daily basis.

In this paper, we introduce a novel method to resolve conflicts for a given field value in relation to the assessed reliability of the collections that contribute data to ASTRIAGraph. ASTRIAGraph, is implemented in a Neo4J graph database [2]. In the database each field from a collection and its corresponding values are viewed as nodes. We treat nodes within collections and common nodes across collections as connections from one node to another, and we have adapted the PageRank algorithm to shift weights around the graph according to these connections [3]. The methods of setting and distributing weights are based on data curation best practices. The end result is a Weight Distribution Network (WDN) that provides a curation-based computational assessment of the collections' reliability and can suggest resolutions to differences (conflicts, gaps, overlaps) between the different collections' data values.

3. WDN INFORMED BY DATA CURATION

Data curation entails activities towards maintaining, preserving, and assuring the quality of research data throughout the continuum of generation and reuse [4]. ASTRIAGraph has complex curation challenges in that it aggregates data about ASOs from different collections into a single knowledge system. These collections contribute similar and different static and dynamic data fields and each is updated at a different pace. This team has worked on two data curation projects for ASTRIAGraph. In [2] we developed a unified, expandable, and scalable data model implemented to display the provenance of data recovered across multiple data sources as a result of scientific queries. Building on that, the Synchronic Curation framework [5] includes a collection analysis and comparison module to track updates and to identify gaps, changes, and irregularities within and across collections. Building on these research, in this work we take a step further and address an ASO resolution. Currently ASTRIAGraph displays individual ASOs' data provided by each collection, rendering a view of multiple repeated ASOs. Given similar field values contributed for a same ASO and considering gaps, changes and irregularities across collections, our system can automatically suggest which collection provides the best field. An important concept in data curation is reliability, which is the confidence that can be placed on the data provided by a given collection. Translated as different aspects of a collection's quality, reliability is assessed by curators using established data curation best practices. In this work we assess collections' reliability through three metrics: a) *completeness,* b) *coincidence*, and c) *consistency*. The metrics are established as punishments or rewards using an adaptation of the PageRank algorithm to identify, for a given data field conflict, which one is the most reliable collection providing that value.

4. RELATED WORK

We review three systems with a similar motivation as ours, data resolution. Data tamer offers a system that uses AI to merge collections with different formats and field names [6]. Merging occurs when the system detects that collections reference the same object, and this subprocess covers gaps by including fields from the multiple collections. During a final step, it automatically flags cases where human input is needed to resolve a data conflict. This system does not offer total resolution, but it manages the composite system to reduce the order of resolution cases.

SLiMFast is another system to resolve inconsistencies between data sources by treating data fusion as a statistical learning problem [7]. Rather than collections, the inputs are individual research articles, and the system tries to extract gene and disease data. Based on the input rows, along with its available pool of previously processed data, the system makes assessments of source reliability. Users of SLiMFast can also label certain entries as truth. The system then computes the accuracy of the input data with a probabilistic model. SLiMFast is a flexible hybrid of domain-specific user input and a statistical estimate of reliability.

Reference [8] defines a system to rectify errors in published chemogenomics research. First it curates chemical data by correcting compound inaccuracies, which are based on known chemical processes. Next, it looks to match bioprofiles, relying on experiments involving the same sets of chemicals. These two steps provide a framework to examine collections from different labs and gain a sense of collection reliability. Entries in the collections are verified by experiments, so there is an assumption that entries treated as truth can be verified if needed.

The three systems reviewed are concerned with data resolution and their assessments are partially based, or finally resolved through human feedback or experiment validation. Our project differs from these in that collection reliability is purely based on metrics of curation best practices implemented in an algorithm that distributes weights according to how these best practices are evidenced in each collection and across collections.

5. METHODOLOGY

Originally based on citation metrics, PageRank is an algorithm used by Google to rank the importance of websites for optimizing search results [3]. The algorithm represents a set of worldwide web pages as nodes in a graph, and the hyperlinks embedded in web pages create the links between nodes. The importance of each page is set as a weight, and weight is distributed along links. Web pages which are referenced more will accrue more weight. The end result is a ranking of importance, driving the order of returned search results, derived only from the content and structure of website data. The PR algorithm was defined in a generic form to be applicable to any linked database of documents that reference each other.

To illustrate the capability of the WDN system, in Table 1 we show the characteristics and scope of the 7 collections that provide data about ASOs to ASTRIAGraph. The names of the collections have been anonymized and replaced with a collection ID. The table shows, per collection, the number of ASO's and corresponding fields (nodes) that they report on as well as the frequency with which the data is updated.

Note that not all field values may change at every update. Data provided by the collections can be categorized as static and dynamic. Static fields contain the same values throughout different data updates (e.g. Country or NORAD ID), while dynamic fields (e.g. eccentricity) contain values that change during updates across time. The assessment is facilitated by our previous work devising a data model through which we normalized the field labels and their definitions across collections [2].

Table 1: Scope of Collections Providing Data About ASOs to ASTRIAGraph

Collection ID	Number of ASOs	Number of Fields/Nodes	Collection Update Frequency
0	23983	20	daily
1	352	18	daily
3	7081	16	daily
4	210	18	daily
7	4542	15	semi-monthly
13	669	13	single import
15	733	17	single import

Adaptation of the PR algorithm in the ASTRIAGraph Neo4J database is as follows. A PR node is defined as a collection's field linked to related nodes encompassing the field's current and historical data. Initial weights of nodes are set according to curation metrics that indicate the quality of that node's data. Here, instead of passing weight according to hyperlinks, we consider two kinds of relationships through which weights are distributed: a) relationship between nodes within the same collection and b) relationship between similar nodes in different collections. Our previous work normalizing the fields of all the collections into a unified data model enables this PR adaptation [2].

Any node A will have links pointing to other nodes. Some links will point within the same collection. A link will point outside the collection if there is a similar node in another collection. The weight of A is split across these links and distributed to the destination nodes, but the partitioning of weight to links is not necessarily equal from between the links. When the graph is initialized, links are given a scaling factor variable which determines how much weight is passed from the origin node onto the link. Consider node A having weight w and set of links $L = \{l_1 .. l_m\}$ to other nodes. Each link l_i has a scaling factor f_i such that $0 \leq f_i \leq 1$. The weight distributed to link l_i, here labeled w_i, is the total weight w scaled by the ratio of f_i to the sum of all factors in L, or

$$w_i = w \times \frac{f_i}{\sum_{j=1}^{m} f_j}. \quad \textbf{(1)}$$

This method ensures that weight is split proportionally between all available links. If node A's definition states that it is related to a node B in the same collection, then the link from A to B receives a scaling factor of 1. For links across collections, our coincidence metric assigns the scaling factors. All other links receive a scaling factor of .05 which ensures that all nodes will send and receive some weight.

The three curation metrics - completeness, coincidence, and consistency - determine the initial weights of the nodes and the scaling factors of links across collections. These metrics represent patterns of gaps, changes, conflicts, and commonalities in data values within and across collections.

Completeness evaluates the presence of gaps within the most recent data in a collection's node. Gaps can be blank values or strings denoting an unknown value such as "TBD." This metric outputs the percentage of filled values in the node's most recent data as a value between 0 and 1.

$$T(A) \text{ is the total number of current entries in node } A \quad \textbf{(2)}$$

$$G(A) \text{ is the number gaps in node } A\text{'s current data} \quad \textbf{(3)}$$

$$c_m(A) = 1 - \frac{G(A)}{T(A)} \text{ is the resultant completeness score} \quad \textbf{(4)}$$

Coincidence compares the current data values between two similar nodes in different collections. It returns the percentage of values in one node's present in the other node's data. We denote the coincidence of node A with B as $c_i(A, B)$, and similarly of node B with A as $c_i(B, A)$. We have

$$N(S) \text{ is the number of values in set } S \quad \textbf{(5)}$$

$$I = A \cap B \quad \textbf{(6)}$$

$$c_i(A, B) = \frac{N(I)}{N(A)} \text{ and } c_i(B, A) = \frac{N(I)}{N(B)}. \quad \textbf{(7)}$$

As a percentage, coincidence is bounded between 0 and 1.

Consistency considers the historical data in a single node. By historical data, we refer to data that aggregates over time through updates noted in Table 1.We chose, for simplicity, to sample each node over the span of a year with steps of 10 days. Between concurrent samples, we look for changes, whether new gaps or altered values. As such, this metric is designed to be completeness and coincidence within the historical data of a single node. The full definition is as follows:

$$N(A) \text{ is the size of some set } A \quad \textbf{(8)}$$

$$S \text{ is the set of historical samples} \quad \textbf{(9)}$$

$$Y(j) = \frac{N(I)}{N(S_j)} \text{ where } I = S_j \cap S_{j+1} \quad \textbf{(10)}$$

$$c_n = \sum_{j=1}^{N(A)} \frac{Y(j)}{N(S)} \text{ is the resultant consistency score.} \quad \textbf{(11)}$$

Consistency is bounded between 0 and 1, as it is a normalized sum of percentages.

We use completeness and consistency to initialize the weights of the nodes. Simply, completeness plus consistency gives the starting weight of a node. This bounds the starting values between 0 and 2. As shown above, equation 7 provides the coincidence between nodes A and node B. We assign the result of equation 7 as the scaling factors of the links between node B to node A. Coincidence computes how much of one node's data is a subset of the other, and larger coincidence increases the amount of weight to be distributed.

After the graph is initialized, we begin a weight distribution loop that considers one node at a time. First, the new weight of each node is computed as the sum of all weights on links pointing toward this node. The new weight is compared to the previous weight; if the difference is less than a threshold, here chosen as 1×10^{-16}, then the node's weight has converged. Finally, weight is distributed onto links pointed away from the node. The portion of weight sent to each link is determined by the scaling factors of links, according to equation 1, as described above. This process continues until the weights of all nodes converge within the same iteration. We take the final weights as the reliabilities of the nodes. This provides a basis for evaluating the quality of the data collections field-by-field. This system allows making an informed choice between collections when field values do not match. If two values agree, then we would consider the reliability behind that value the sum of the weight of the two nodes. This summed reliability could be compared to the reliability of a node with a different value.

The weight distribution process converges in under 1000 loops around the graph. The greatest changes in weight occur within the initial 25 iterations, followed by a smooth convergence of the weights for the remainder of the process, as shown in Fig. 1 where each line represents the changing weight of a node throughout the weight distribution process.

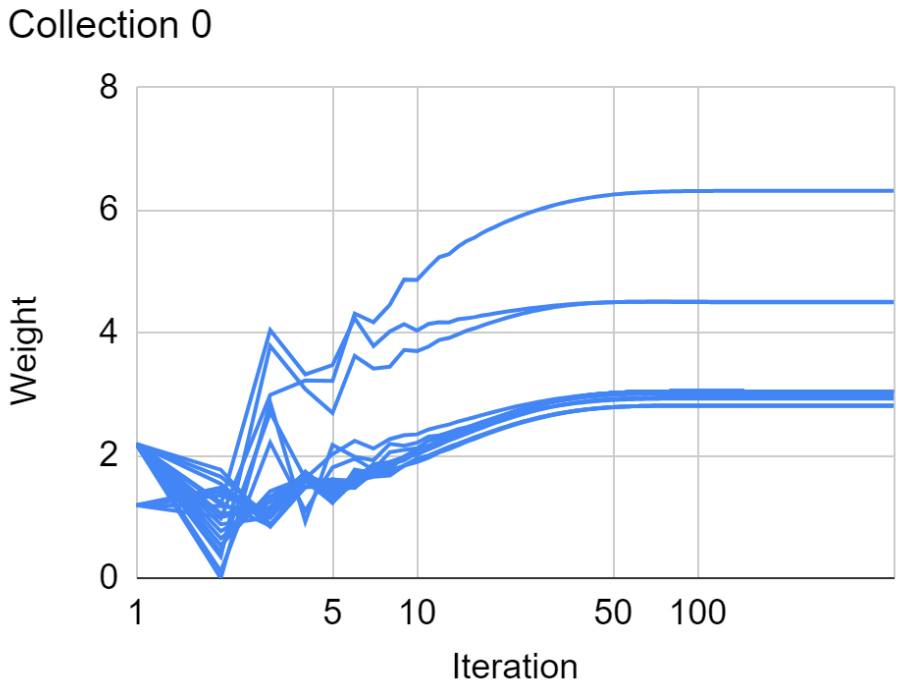

Fig. 1: Graph showing node weight iteration for one of the collections Collection 0

Fig. 1 shows the node weight distribution process of all nodes in collection 0. This collection regularly delivers a high volume of data tracking the greatest number of ASOs in ASTRIAGraph's set of collections. Moreover, most of its nodes contain complete values. In the graph we can observe that some nodes are more reliable than others, but we found that the overall final weights for the collection are amongst the highest.

6. USE CASES RESULTS

We illustrate how WDE delivers final results through use cases involving data conflicts resolution of static and dynamic fields between three collections.

The first case involves resolving for a particular ASO, if collection 0, 1, or 7, is the most reliable to provide the field value "Country." As shown in Table 2, only collection 0 contains a small percentage of gaps in its current data, and the three scores are very similar.

Table 2: Country Completeness Scores Between Collections 0, 1, and 7

Node Name	Completeness
0_Country	0.989
1_Country	1.0
7_Country	1.0

Table 3 shows that the consistency scores are also similar, with only collection 7 containing changes in its historical data, though also a small number.

Table 3: Country Consistency Scores Between Collections 0, 1, and 7

Node Name	Consistency
0_Country	1.0
1_Country	1.0
7_Country	0.991

The last metric, coincidence, affects this case the most. Results shown in Table 4 indicate that Collection 0 contains almost all of the Country values present in 1 and 7, while Collection 1 contains almost no values from the other two. The completeness and consistency scores initialize the three Country nodes with roughly the same weight, but the coincidence scores funnel large portions of weight from 1_Country and 7_Country into 0_Country during the weight distribution process.

Table 4: Country Coincidence Between Collections 0, 1, and 7

Compared Nodes/Fields	Coincidence
0_Country with 1_Country	9.901×10^{-3}
0_Country with 7_Country	9.901×10^{-3}
1_Country with 0_Country	1.0
1_Country with 7_Country	0.0
7_Country with 0_Country	1.389×10^{-2}
7_Country with 1_Country	0.0

Collection 0 contains data on many objects - the most of our collections. In contrast, collection 1 only tracks some United States objects. Collection 7 contains data on various objects from many countries, but the size of collection 7 is moderate - not small or large. This is why 1_Country and 7_Country are almost entirely subsets of 0_Country.

Table 5: Country Reliability Scores for Collections 0, 1 and 7

Node/Field Name	Reliability
0_Country	2.928
1_Country	0.190
7_Country	1.186

Overall, Collection 0 is found to be the most reliable to resolve the field Country. Reliability ranks for the three collections are shown in Table 5, and the convergence of the weights is shown in Fig. 2. The coincidence values, which most strongly affect the final reliability scores in this case, are subject to the characteristics of the collection to which the values belong, which our metric measures. The domain experts in our team, who have purview on the sources of the data collections involved in the study, agreed with the results obtained, and indicated that Collection 0

is a renown, established collection, generated and used by a trustworthy and diligent organization with capable equipment.

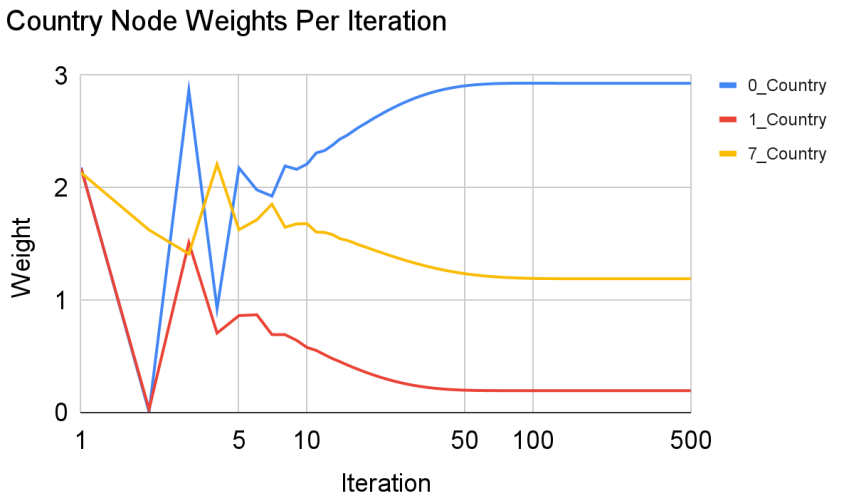

Fig. 2: Convergence of Country Node Weights Reliability in Collections 0, 1, and 7

The next case considers the resolution of dynamic values. In Table 6 we present an example with two collections (0, 1) providing eccentricity values for the same ASO. Eccentricity defines the shape of an ASO's orbit. While collections provide a near-0 value, denoting a circular orbit, collection 0 claims a tighter circularity than collection 1. The differences between these values is one order of magnitude, where each of these numbers has 16 decimal digits of precision.

Table 6: Eccentricity Differences Between Collections 0 and 1 for the Same ASO

ASO's NORAD ID	0_Ecc	1_Ecc
41964	9.137×10^{-4}	2.462×10^{-3}

The algorithms that predict orbital states, which vary between collections, are sensitive to the high precision of their variables. In turn, inputs, whether previous orbital states, radar or telescope measurements, or otherwise, are subject to the time of the tracking measurements. The WDE assessment neither has knowledge of the specifics of the algorithms and tracking equipment, nor examines the differences in time of data generation that characterizes dynamic data. Purely based on curation metrics, we find that Collection 0 has higher reliability than collection 1. A more robust collection reliability assessment would require building in a domain specific analysis of eccentricity that considers numerical fields with high precision and includes time of data updates in the comparison.

Table 7: NORAD ID reliability scores.

Node Name	Reliability
0_NoradId	4.509
1_NoradId	0.258
3_NoradId	0.998
4_NoradId	0.581

Curation-based WDN can inform more complex SDA decisions. For example, occasionally orbital states from different database entries will coincide; that is, sets of elements or state vectors will be similar, but the associated identifiers will not match. This can be caused by an error in the tracking systems, intentional mission design of a rendezvous trajectory, or error within the collection(s) alone. Our work offers resolution of the ASO identifiers tied to the orbital data. Using the reliability computed for NORAD ID nodes, we would put more trust in the identifiers of collection 0, followed by collection 3, as shown in Table 7. The data in these collections could indicate whether similar orbits are duplicates for one object or multiples close together. These types of insights could be used to augment the work of experts in orbital determination and object tracking.

7. CONCLUSIONS

We developed WDN to identify the reliability of collections providing data to the ASTRIAGraph SDA knowledge system. We showed that the method is useful to resolve conflicting values by helping identify the most reliable collection providing the data. Based on an adaptation of the PR algorithm informed by three curation metrics, the method is particularly useful to resolve static data conflicts and we understood the limitations when evaluating dynamic fields. Indeed, comparison operations are non-trivial for the dynamic/numeric fields in our collections. Future requirements to augment this system call for domain-specific algorithms to analyze the similarity of dynamic variables programmatically. Additionally, while this work considers each field within a collection individually, some fields come in sets. For example, eccentricity, discussed above, is of course one variable in a set of six orbital elements, and these elements can be derived from and be used to derive an ASO's 6x1 state vector of position and velocity. Also, definitions of some fields are partially related, for example CosparId and BirthDate. We posit that these relationships could be modeled with intermediate nodes added to the graph, indicating the information embedded in the data. Weight could then be distributed between variable sets as a whole. The main contribution of this methodology is that it addresses the reality of large-scale, multivariate, multi-collections data curation automatically and establishes a valid path towards achieving transparency and improving the quality of data-driven science.

8. REFERENCES

1. "ASTRIAGraph." *The University of Texas at Austin*. http://astria.tacc.utexas.edu/AstriaGraph/. Accessed Aug. 2023.
2. Esteva, Maria, et al. "Modeling Data Curation to Scientific Inquiry: A Case Study for Multimodal Data Integration." *Joint Conference on Digital Libraries*, 2020.
3. Page, Lawrence. "Method for Node Ranking in a Linked Database." US 7058628 B1, *US Patent and Trademark Office*, 7 June 2006.
4. Pouchard, Line. "Revisiting the Data Lifecycle with Big Data Curation." *International Journal of Digital Curation*, vol. 10, no. 2, June 2015, pp. 176–92, 2015.
5. Esteva, Maria, et al. "Synchronic Curation for Assessing Reuse and Integration Fitness of Multiple Data Collections." *International Journal of Digital Curation*, 2022.
6. Stonebraker, Theodoros, et al. "Slimfast: Guaranteed Results for Data Fusion and Source Reliability." *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017.
7. Fourches, Michael, et al. "Data Curation at Scale: The Data Tamer System." *Conference on Innovative Data Systems Research* Vol. 2013, 2013.
8. Rekatsinas, Theodoros, Eugene Muratov Denis, and Alexander Tropsha. "Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation." *Journal of Chemical Information and Modeling* 56.7, 1243-1252, 2016.