Visualization of Massive Movement Datasets with Low Barrier to Entry – A Case Study

Alexander Savelyev
Department of Geography
Texas State University
San Marcos, TX, USA
savelyev@txstate.edu

Abstract — With large movement datasets commonplace, a special focus in the information visualization community emerged with the goal of developing theories, frameworks and techniques tailored to the challenges and opportunities associated with the analysis of movement data. Looking to accommodate the adoption of methods and techniques of the emergent movement visualization science across a range of research communities, a call is made to look for balance between the sophistication and power and the steep learning curve associated with the production and use of advanced visualization tools. This paper presents a proof-of-concept approach to the exploratory visualization of massive movement datasets using well-established open-source software tools designed specifically for the non-programmers. This approach is evaluated by means of a case study — an exploratory visualization of 120 million movement records derived from social media data — that yields actionable insights for an analyst looking to explore, clean, and organize the sample dataset, or to calibrate and parameterize theoretical models for its subsequent in-depth analysis.

Keywords — movement, visualization, geovisual analytics, BigData

I. Introduction

Large movement datasets are now fairly commonplace — examples include trajectories collected by means of explicit GPS tracking of animate and inanimate entities, by means of processing toll data in various transit systems, analysis of various state and municipality datasets, as well as volunteered movement data of various kinds, such as geotagged social media. The sheer size of such datasets and the frequent ease with which they are compiled made their analysis a priority and have contributed to a surge in interest in data-driven (as opposed to hypothesis-driven) research [1].

The theories and methods of information visualization and (geo)visual analytics are currently understood to play a critical role in the emerging integrated "science of movement", as they make it possible to engage in data exploration, data-driven ideation and hypothesis generation [2], data cleaning and pre-processing, calibration and parameterization of computational models [3], as well as data analysis proper. The visual approach to geographic data analysis has a further advantage in that it makes possible for the analyst to take advantage of their implicit understanding of spatial relationships, spatial context, and other critical insights that are currently impractical to model using a computational approach alone [4].

Visualization of large datasets, however, is hard. From the technical standpoint, such datasets routinely exceed the size of available computer memory and pose a formidable challenge for the design of algorithms that could process them in a reasonable amount of time [5]. From the visualization design standpoint, such datasets routinely cause excessive overplotting, rendering the resulting charts indecipherable [4]. In response to these challenges, a wide range of clever prospective solutions has been put forth, including various methods for managing the visual complexity of the displays [6, 7] as well as interactive visualization techniques that scale to millions and billions of records [5, 8].

Over time, these efforts saw great adoption across a wide variety of fields that share an interest in movement visualization. Following this growing interest, a new call emerged for an attempt at a balance between the *sophistication and power* of the methods developed in the information visualization and (geo)visual analytics communities, and the *steep learning curve* associated with their adoption in other fields [1, 3].

This paper aims to demonstrate that such balance is possible. Using well-established open-source software tools, designed specifically for the non-programmers, this study describes a proof-of-concept approach to the exploratory visualization of massive movement datasets. This approach is further evaluated by means of a case study — an exploratory visualization of approximately 120 million movement records derived from geotagged Twitter data — that yields actionable insights for an analyst looking to explore, clean, and organize the sample dataset, or to calibrate and parameterize theoretical models for its in-depth analysis.

II. Methodology

One of the software components used in this study is *Processing* (https://processing.org/) — a highly-simplified, open-sourced graphical library and integrated development environment, designed to introduce artists, designers and other "non-programmers" to building interactive information visualizations through code. *Processing* makes use of a raster data model (a large drawing "canvas" composed of individual pixels) and a simple loop that runs user-specified code at a certain frequency (e.g. 30 times per second), imitating animation frames. *Processing* also provides a wide assortment of "commands" such as *point()*, *line()* or *arc()* that reduce the

complexity of drawing on the screen to invoking a single command with pixel coordinates as input.

Besides *Processing*, this proof-of-concept methodology makes use of two further items — the structure of the movement data file and the movement visualization algorithm.

The movement data file contains records in the format of "coordinates of origin — coordinates of destination — time of departure — time of arrival — moving object ID". Each record corresponds to a segment of a movement trajectory for a particular object or person. The movement file is sorted chronologically, earlierst records first, using the time of arrival to perform the sorting.

The movement visualization algorithm is, in essence, a form of movement animation, and applies the following logic to each record in the data file, in order:

- If the timestamp of the movement record fits in the current animation frame,
    - if this movement record satisfies filtering parameters,
        - project the movement record coordinates to the screen coordinates,
        - select an appropriate visual style for the movement record, and
        - draw the movement record as a straight line.
    - else,
        - skip this record.
- else,
    - fill the screen with translucent black to imitate "fading" of older records, and
    - advance to the next animation frame.

In this proof-of-concept, segment length and duration were used as the filtering parameters in the algorithm outlined above, but many other metrics are available [9]. This proof-of-concept employed a variant of the stereographic projection, but raw coordinates or any other transformation could be used. Similarly, the visual style employed could be as simple as a solid white line of a given thickness, or determined as a function of the movement record data. The size of the animation frame (e.g. 1 minute of real time per frame) and the intensity of the translucent black fill (e.g. 10%) are selected arbitrarily to manipulate the speed of the animation and the speed with which older records "fade away".

Since this algorithm works on a single record at a time and does not store any past state (beyond that captured by the raster drawing surface), it has no restrictions in terms of the file size, and can be applied to the visualization of infinite, streaming datasets, in both desktop and online environments. The filtering criteria are applied on a per-record basis as well, which makes it trivial to adjust the filtering mechanism on the fly (see the accompanying study video — https://youtu.be/gWoiTI_5O_U — for an illustration of interactive filtering controls at work).

Following this proof-of-concept methodology, the study author put together a *Processing* app and made a series of video recordings, with different filtering parameters and different aspects of the projection (center point and zoom level) employed. The resulting recordings were then reviewed, with the author assuming the role of a data analyst. Some of the most salient insights obtained in this process are illustrated in the study video and described in detail in the sections below.

Three sample datasets, formatted as described earlier, were used in this case study. First dataset contains movement records derived from the 80 million geotagged Tweets sent anywhere in the western hemisphere in the December of 2017 (using 95 to 99% of all geotagged tweets sent during that period). The second and third datasets were produced in a similar fashion, but contain data from April of 2019 (30 million) and 2020 (10 million), respectively.

III. Results

This section is structured to reflect the various strengths attributed to the visual approach to data analysis in the *Introduction* section, specifically its ability to support data exploration and ideation processes, its contribution to the data cleaning and pre-processing steps, and its potential role in the selection, calibration and parameterization of advanced computational models.

A. Data exploration and ideation

Some of the most visually salient aspects of the resulting visualization are those related to data availability and volume. For example, as seen throughout the study video (in particular at timestamps 1:10 – 4:40), there are clear distinctions between data volume by country and by region, and clear rhythms to the data volume generated during different parts of the day. A clear, persistent correlation between the temporal and spatial patterns is also present, as the morning spikes in data volume seem to overlap well with the geographic borders of different time zones.

Placing two instances of the visualization app, it is also possible to make an attempt at side-by-side comparison. Although simultaneous animation is likely a suboptimal technique for such analysis, it readily highlights a striking difference in data volume (video timestamp 4:40) in the Washington, DC – Boston corridor between April of 2019 and 2020 (the rough date of COVID-19 lockdown restrictions coming in effect in the said area). Comparatively speaking, the Christmas travel rush is much less pronounced when comparing travel on Friday, December 22 to that a week prior (timestamp 5:40).

B. Data cleaning and pre-processing

If the filtering parameters in the app are adjusted to focus on the *improbable*, and the speed with which records fade is reduced to let data accumulate, an image similar to that shown in Fig. 1 can be obtained. This figure shows the movement segments corresponding to entities travelling farther than 2,500 km (1,500 mi) in less than a minute — no small feat! A brief inspection of the tweets corresponding to these movement

records suggests bot activity, which is likely a valuable insight for the hypothetical data cleaning protocol.

C. Computational model selection and parameterization

Another visually salient aspect of the resulting visualization is the general absence of obvious trajectories. Most movement records appear as segments that are either disjoint (i.e. belong to different entities) or are too far apart in time to see a trajectory appear using the visualization metaphor employed. The few trajectories that do appear, however, are quite striking, often covering large distances or even appearing as traces of a regular "commute" (timestamp 6:40). Either observation could be described as a criteria for the model selection. For example, bulk generalization of detailed trajectories is likely impractical with this dataset due to the sampling sparsity (few records per entity). On the other hand, detailed trajectories clearly exist, but it is unclear what specific motivation triggers their production by the moving entities.

When rendered using larger cartographic scale, further insights are possible. For example, Fig. 2 and Fig. 3 show the same region (Manhattan, NY), but Fig. 2 is restricted to showing movement segments under 1.5 km (1 mi) long. In both figures, clear "anchor" locations are visible, likely corresponding to place centroids (e.g. those of neighborhoods), but they appear less dominant in Fig. 2. Again, this observation could be described as a criterion for the prospective model selection. For example, if fine spatial resolution is needed, it might be best to filter the data accordingly to avoid over-representation of place centroids in the model results.

IV. Discussion

This study set out to demonstrate that it might be possible to strike a balance between the complexity of the visualization methods and tools and their power, with movement visualization as a specific case study. As elaborated in the Introduction section, it is not uncommon for the practitioners of the emerging science of movement to have markedly different backgrounds and technological skillsets, whereas modern movement datasets are often large and present a formidable computational visualization challenge. While plenty of cartographic tools are available (ranging from desktop solutions such as ArcGIS Pro to web-based applications and APIs such as Leaflet or Mapbox), they do not provide ready and scalable solutions for large movement data visualization. On the other hand, while plenty of clever engineering and theoretical solutions to BigData visualization problems are available in the literature, they are seldom of much use to scholars and practitioners who do not possess the requisite technical background to implement those solutions on their own. This study demonstrates that it is, however, possible to conceptualize a BigData visualization approach with low barrier to entry by means of carefully picking the data models (a raster drawing surface and a chronologically sorted movement data file) and the algorithms employed (processing data records sequentially and performing data filtering on the fly).

This proof-of-concept approach appears to perform adequately for the purpose of an exploratory analysis of a sample large movement dataset. Although in no way comprehensive, the insights drawn from this analysis would likely be sufficient to have an impact on the process of ideation, hypothesis generation, data cleaning, and computational model selection, if these were sought by the hypothetical analyst.

Despite a certain measure of merit, however, this proof-of-concept methodology also has a number of potential drawbacks that might limit its utility. For example, this methodology relies on a form of movement animation to work around the technical issues of handling very large datasets. The efficacy of animation as a data analysis tool, however, is currently poorly understood [7, 10], and the insights highlighted above only speak to what could be seen, but not to what was missed. Another shortcoming of this approach is that, when applied to datasets of actual infinite size, it effectively becomes a transient visualization — there is no possibility to "rewind" or otherwise review the past frames. Yet another potential issue is that the key premise of this paper – that the approach chosen has a comparatively low barrier to entry – could be better supported by means of an empirical user study. Given that the hypothetical tasks supported by this approach (ideation, hypothesis generation, analytical model conceptualization) fall within well-established geovisual analytics scenarios [11], such user study would allow to directly comment on the usability and utility of this approach, along with its contribution to the overall geovisual sensemaking process.

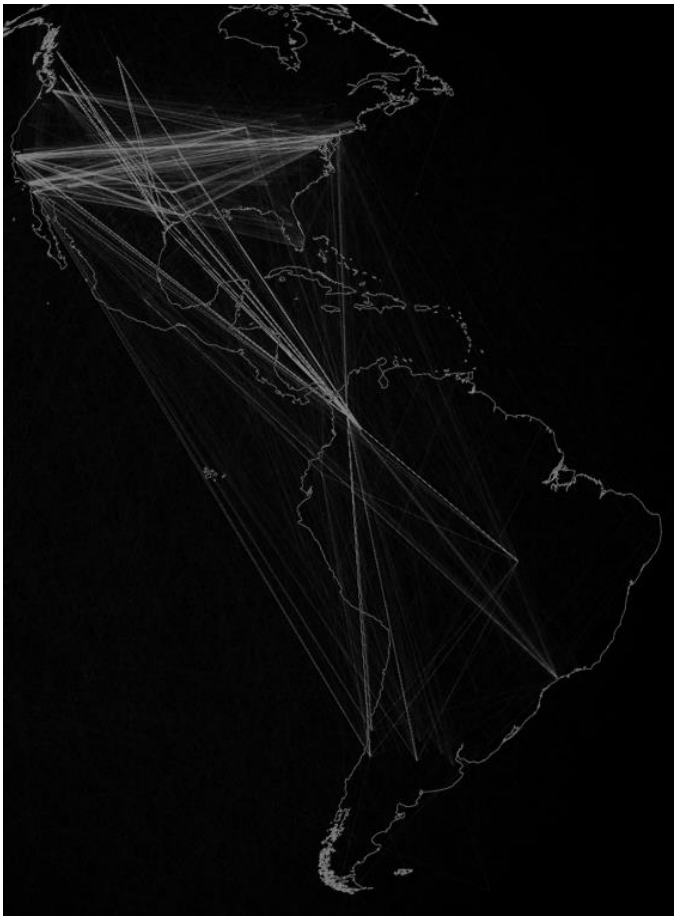

Fig. 1. Movement records across North and South America, filtered to highlight travel over 2,500 km (1,500 mi) in less than a minute.

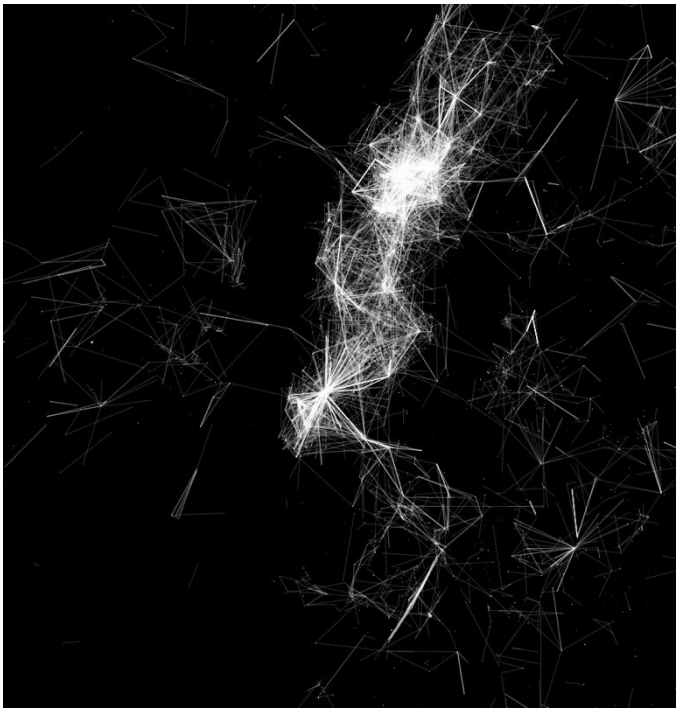

Fig. 2. Movement records across Manhattan, NY, filtered to only show the segments that are less that 1.5 km (1 mi) long.

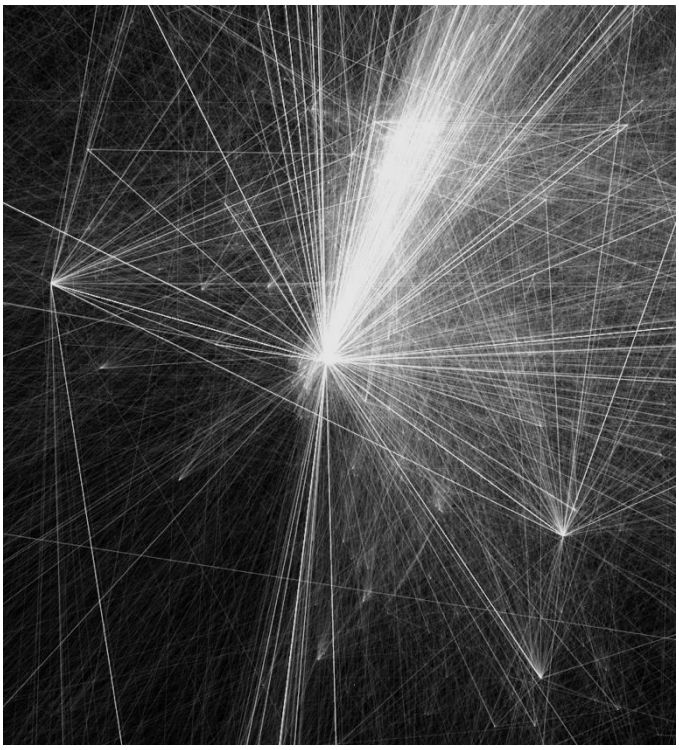

Fig. 3. Movement records across Manhattan, NY, with no filtering applied.

REFERENCES

[1] M. Thums, J. Fernández-Gracia, A.M. Sequeira, V.M. Eguíluz, C.M. Duarte, and M.G. Meekan, How big data fast tracked human mobility research and the lessons for animal movement ecology. Frontiers in Marine Science, 5, p.21., 2018

[2] S. Dodge, A Data Science Framework for Movement. Geographical Analysis, 2019.

[3] H.J. Miller, S. Dodge, J. Miller, and G. Bohrer, Towards an integrated science of movement: converging research on animal movement ecology and human mobility science. International Journal of Geographical Information Science, 33(5), pp.855-876, 2019.

[4] G. Andrienko, N. Andrienko, J. Dykes, S.I. Fabrikant, and M. Wachowicz, Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research, 2008.

[5] Z. Liu, B. Jiang, and J. Heer, imMens: Real‐time visual querying of big data. In Computer Graphics Forum (Vol. 32, No. 3pt4, pp. 421-430). Oxford, UK: Blackwell Publishing Ltd., 2013.

[6] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, and Y. Zhao, Visual analytics of mobility and transportation: State of the art and further research directions. IEEE Transactions on Intelligent Transportation Systems, 18(8), pp.2232-2249., 2017.

[7] A. Çöltekin, A.L. Griffin, A. Slingsby, A.C. Robinson, S. Christophe, V. Rautenbach, M. Chen, C. Pettit, and A. Klippel, Geospatial Information Visualization and Extended Reality Displays. In Manual of Digital Earth (pp. 229-277). Springer, Singapore, 2020.

[8] K. Soltani, A. Padmanabhan, and S. Wang, MovePattern: Interactive framework to provide scalable visualization of movement patterns. In Proceedings of the 8th ACM SIGSPATIAL International Workshop on Computational Transportation Science (pp. 31-36), 2015.

[9] W.R Tobler, Experiments in migration mapping by computer. The American Cartographer, 14(2), pp.155-163, 1987.

[10] M. Scaife and Y. Rogers, External cognition: how do graphical representations work?. International journal of human-computer studies, 45(2), pp.185-213. Vancouver, 1996.

[11] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. IEEE transactions on visualization and computer graphics 18(9), pp.1520-1536, 2011.