Journal of Maps, 2013
Vol. 9, No. 4, 481–486, http://dx.doi.org/10.1080/17445647.2013.824391

SOCIAL SCIENCE

Exploring geo-genealogy using internet surname search histories

Yifan Zhang[a], Muhammad Adnan[b*], Paul Longley[b] and Ross Maciejewski[a]

[a]*School of Computing, Informatics & Decision Systems Engineering, Arizona State University, Arizona;*
[b]*Department of Geography, University College London, London*

(*Received 6 March 2013; resubmitted 3 June 2013; accepted 24 June 2013*)

We present an interactive flow map to visualize aspects of the ways in which surnames have dispersed and migrated around the globe. This work utilizes Internet search queries from the Worldnames Project and uses the density of search locations to determine the node and leaf structures of a flow map. The mapping technique utilized in this work is a variant of geometric minimal Steiner arborescences called the *spiral tree*. Our implementation is developed in JavaScript to allow for interactive online exploration. Nodes and flow lines can be interactively modified to allow for esthetic changes of color and layout. The results can provide interesting insight into the geography of amateur genealogy.

Keywords: surname; genealogy; flow map

1. Introduction

The study of family lineage and history is a well-established field of research, both amateur and professional, that uses historical records and genetic analysis to ascertain kinship. With the advent of digital encoding of historical records, and their wide availability through the Internet, more and more amateur genealogists are able to explore the global reach of their family lineage. A related innovation is that of 'geo-genealogy', in which the geographic origins and contemporary spatial distributions of surnames can be ascertained and mapped. For example, work by Cheshire et al. (2010) has explored the regional basis to surname distributions in Great Britain, using part of a database of names that is becoming increasingly global in coverage (http://gbnames.publicprofiler.org).

The focus of this map is the database that underpins the Worldnames website (worldnames.publicprofiler.org), which provides users with online maps of surname distributions at the global scale and a range of regional levels. Since its launch in August 2008, this website has been visited by more than 3.1 million unique users. In order to better measure the outreach and impact of the site, the location of the ISP address for every new search, along with the name searched for, has been recorded since July 2011. As of February 2013, a total of 318,838 paired names and locations have been stored.

The map accompanying this commentary extends the Worldnames mapping application by representing names searches as interactive flow maps. The information arising from name

*Corresponding author. Email: m.adnan@ucl.ac.uk searches is likely to have been stimulated by historic migration of bearers of the searched for surnames across the globe. Users may access the online version of the map at http://www.uncertaintyofidentity.com/SurnameSearch.html in order to explore potential migration patterns of all surnames that have been searched for, in addition to those shown in our map. The results of surname searches are presented as flow maps in which the root of the flow identifies the highest share of the local population accounted for by the selected surname in any of the 26 countries for which surname data are available on the Worldnames site. These name 'origins' are represented as points as a way of summarizing the local and regional distribution of the name. In many cases, particularly when a name is rare and has a unique origin, this will identify the approximate location at which the name was first coined. In other cases, more common names (such as Smith or Brown) had multiple origins, and the point locations will thus summarize a wider distribution rather than providing an accurate point of origin. In such cases, the point location nevertheless provides an indication of the broader area in which a name originated. Further details on the origins of many established Anglo Saxon names in Great Britain may be found in the data pages of http://gbnames.publicprofiler.org, and similar information for other countries can be obtained from the tables accompanying the maps generated at http://worldnames.publicprofiler.org. The maps provide an indication of the Diaspora of many surnames and can stimulate users to suggest hypotheses about how their own family's history corresponds with the global pattern of searches.

Flow maps combine maps and flow charts as a means of showing the movement of objects from one location to another, making them ideal for communicating surname migration patterns from the location (or region) in which the name was first coined to the locations at which present day bearers of the name query the Worldnames database. One of the earliest examples of a flow map was Minard's map of French Wine exports in 1864 (Minard, 1864). Such maps proved informative and esthetically pleasing; however, the time needed draw them and the complexity in distributing flow lines often made the creation of such maps intractable. However, in recent years, thanks to increases in computational power, researchers have begun developing interactive computer-based tools for creating flow maps.

One of the earliest tools was Flow Mapper (Tobler, 1987, 2003) which allowed for the production of a total movement map shown by volume-scaled bands. Later work by Dodge and Kitchen (2004, 2007) utilized flow maps to explore the Internet infrastructures, and work by Cox, Eick, and He (1996) and Munzner, Hoffman, Claffy, and Fenner (1996) extended the concept of flow maps into 3D space. Guo (2009) developed an integrated interactive visualization framework utilizing flow maps to explore multivariate data, and Boyandin, Bertini, and Lalanne (2010) developed a tool for visualizing migration flows over time using animation.

Central to this approach is the development of algorithms for distributing flow lines in 2D geographic space such that they avoid (whenever possible) intersection and are esthetically pleasing. Work by Phan et al. (2005) developed an algorithm for defining the layout of flow maps utilizing hierarchical structures within the data. Cui et al. (2008) utilized control mesh methods for flow line layouts, and Holten and Van Wijk (2009) proposed a force-directed method for bundling edges. Our work utilizes the spiral-tree method introduced by Buchin, Speckmann, and Verbeek (2011) which utilizes *angle-restricted Steiner arborescences* for reducing visual clutter by bundling lines smoothly and avoiding self-intersection.

2. Methods

We have used two datasets for plotting surname searches made on the Worldnames site (http://worldnames.publicprofiler.org). Worldnames is a web service developed under a research project at Department of Geography, University College London, and holds surname data for

26 different countries of the world. These countries include a major part of Europe, USA, Canada, Argentina, India, Australia, and New-Zealand, and account for over a billion of the world's population. On each of the maps, the red dot identifies the location with the highest concentration of the surname in 26 countries that make up the Worldnames database.

Since July 2011, the Worldnames website has been collecting some additional information from users who search for a surname. This is the second database used in this study, which includes the IP addresses of the individuals who conducted the search, the name that they searched for, and the gender of the person conducting the search. Users are made aware that we collect email addresses and locations as evidence of the wide use of the service we provide, and to further our research into the geography of family names. The site terms and conditions also make clear that individual data are not passed on to third parties. Irrespective of gender, it is a reasonable assumption that the overwhelming numbers of users would be conducting searches that are related to their personal family histories. The archived IP addresses are converted to latitude and longitude values by using InfoDB's (http://www.ipinfodb.com/) IP address to Geo-location Application Programming Interface (API). For the maps that accompany this paper we have used the data archived between July 2011 and February, 2013. This database comprises 318,838 entries of surname searches and the corresponding IP address locations of the users. All of the remaining dots on the maps identify the locations at which searches originated for the relevant surname.

In order to create flow maps from these data, we have developed a modified JavaScript implementation of the spiral tree flow map. This method is a recent variant of geometric minimal Steiner arborescences, using a logarithmic spiral tree. The logarithmic spiral segment is an *angle-restricted* path assumed from a point p to its root node r, with the path's self-similar and self-approaching properties being defined by Aichholzer et al. (1998). Specifically, from point p to r, there exists two spirals defined as a right spiral S_p^+ and left spiral S_p^-. These two spirals can be given using parametric equation in polar coordinates assume $p = (R, \phi)$:

$$\text{Right spiral: } R(t) = \text{Re}^{-t},\ \phi(t) = \phi + \tan(\alpha)t$$

$$\text{Left spiral: } R(t) = \text{Re}^{-t},\ \phi(t) = \phi + \tan(-\alpha)t$$

where t is the parameter and α which less than $\pi/2$ is the angle of the spiral, when $0 \leq t \leq \pi \cot(\alpha)$ those two segments will construct a spiral region R_p.

If another point q lies within the region R_p denoted as $q \in R_p$, we will have $R_q \subseteq R_p$ as scaling a logarithmic spiral will result in another logarithmic spiral.

If another point $q \notin R_p$, then there will be an intersection denoted as join point J_{pq} between the spiral segments of p and q, which could be either S_p^+ with S_q^- or S_p^- with S_q^+.

The flow map in our application is then constructed using an enhanced greedy algorithm where each node (or terminal) represents a search for a specific surname on the Worldnames site. For esthetic purposes, we have modified the way in which the spiral tree algorithm creates terminals and joins points. The root of the flow map (colored red) is the location where the highest frequency of a given surname exists within the Worldnames database. An auxiliary circle AC_t is defined as the circle passing terminal t and centered at root r. There is also a set called wavefront W which is a list used for storing the active nodes. It is designed as a balanced binary search tree that organizes active nodes in counter-clockwise radial order around the root, in order to assist the procedure.

The algorithm iteratively joins terminals from the wavefront W until all terminals are connected into a single tree. All of the terminals are initialized as non-active and the wavefront W

is empty. A sweep circle C ranges from the outermost terminals inwards towards the center of root r in order to determine the next node to be processed. During the sweeping procedure, the sweep circle C will encounter two types of nodes: Terminal and Join Point, where a join point is only computed by two adjacent nodes in the wavefront W.

If C reaches a terminal t, t will be noted as an active terminal first and added to W. Then the algorithm examines whether or not t is in any of its neighbor's spiral region W. If there is a spiral region of its neighbor n, then W contains t, and our algorithm finds an auxiliary point x which marks the intersection of AC_t with the farthest spiral segment of n, and then connects x with n using the spiral segment. n is then removed from W and x is treated as a new terminal.

If C reaches a join point J_{pq} and its parent nodes p and q are active in W, then we join its parent nodes p and q to this join point using the spiral segments and remove p and q from W. Next J_{pq} will be noted as active and added to the wavefront W, and the sweeping process continues. At this step we have added a threshold function to position the parent node away from its join point such that if the distance of one parent node p to its join point J_{pq} of q is below a certain threshold, then we will utilize p's auxiliary circle AC_p to find the intersection point x with another spiral segment of node q. After that, q is connected to x using spiral segments and x is taken as a new active terminal point added to W. This is done in order to make terminals more visible to the viewer. In practice, the threshold is set to be 2 to the power of the map zoom level and the default values for the angle of the spiral α is 15 degrees. Users may interactively adjust these properties to create their desired map esthetics.

The JavaScript implementation of the spiral tree flow map is available online at the URL https://github.com/yifantastic/FlowMap.

3. Discussion

As previously stated, maps created using our tool represent the geography of interest in family genealogy, relative to the approximate locations in 26 countries at which different surnames were first coined. For names originating outside these countries, the origin point is defined at the location within the Worldnames' 26 countries at which the highest number of bearers of the name is concentrated. The precision with which the origin locations can be identified varies between names, although this is not an issue at the global scale at which our maps are produced. The maps also provide insights into the geography of interest in genealogy and the likely flows of information between bearers of the same name. This is potentially of use in the marketing of tourist destinations and heritage sites around the world.

Where the flow lines intersect, it is important to be aware that this is the result of the default parameters that are used by the algorithm, or those specified by the user. The curves depicted on the maps do not reflect actual migration flows but rather link a likely historic point in a family tree and the probable end destination of a migrant – ignoring any intervening points in family migration history. The maps may help suggest family migration histories based upon the spatial patterning of locations at which queries were submitted to the Worldnames database.

Our tool also allows users to map multiple flows simultaneously. In such cases, overlap between flows is not considered in the algorithmic layout. However, such overlays can allow users to search for multiple family branches and extrapolate information on potential encounters between family members.

4. Conclusion

This work presents an interactive tool for online exploration of interest in family genealogy. Our maps are illustrations of the outputs of our interactive website at http://www.uncertaintyofidentity.

com/SurnameSearch.html, which allows users to explore the geography of interest in the origins of approximately 130,000 surnames that were the subject of searches on the Worldnames website between July 2011 and February 2013. The branching algorithm that we have adapted presents this information in a clear and uncluttered way, even when large numbers of names searches have been conducted. This is particularly apparent in the online version of this map, which may be viewed at a full range of recursive levels.

Software

The online application that forms the basis to the maps was written in JavaScript. Our software is linked to a back-end database which stores the surname dataset in MySQL and an interactive front-end interface which can be accessed via web-browsers. The data are queried and transferred using AJAX (Asynchronous JavaScript and XML). The flow map representation consists of the background Google map overlain with the SVG network flows.

The Google Map JavaScript API is used for pre-processing the data from the geo-space into the 2D-rendering space and combining the flow layers with the interactive map, available at http://www.uncertaintyofidentity.com/SurnameSearch.html. We also utilized the D3 JavaScript library (http://d3js.org) to generate multiple SVG flow layers and jQuery User Interface library (http://jquery.com) for implementing the user interface. This makes it possible to create multiple tabs for searching different surnames. Two sliders are embedded into the interactive map frame so that users can manipulate the appearance of the flow by adjusting the stroke-width and angle of the spiral tree. Additional information is provided for clickable nodes linked with pop-up windows. Color palette widgets are provided using the JSColor JavaScript library (http://jscolor.com) so that different colors can be applied to distinguish multiple flows.

Acknowledgements

This work was completed as part of the EPSRC research Grant ‘The Uncertainty of Identity: Linking Spatiotemporal Information in the Real and Virtual Worlds’ (EP/J005266/1).

References

Aichholzer, O., Aurenhammer, F., Icking, C., Klein, R., Langetepe, E., & Rote, G. (1998). Generalized self-approaching curves. *Algorithms and Computation*, *Springer*, 317–327. doi: 10.1007/3-540-49381-6_34

Boyandin, I., Bertini, E., & Lalanne, D. (2010). Using flow maps to explore migrations over time, In *Proceedings of Geospatial Visual Analytics Workshop in conjunction with the 13th AGILE International Conference on Geographic Information Science (GeoVA)*, Guimaraes (Portugal).

Buchin, K., Speckmann, B., & Verbeek, K. (2011). Angle-restricted steiner arborescences for flow map layout. In *Abstracts of the 27th European Workshop on Computational Geometry*, Springer, pp. 163–166.

Cheshire, J. A., Longley, P. A., & Singleton, A. D. (2010). The surname regions of Great Britain. *Journal of Maps*, *6*, 401–409. doi: 10.4113/jom.2010.1103. URL http://www.tandfonline.com/doi/abs/10.4113/jom.2010.1103

Cox, K. C., Eick, S. G., & He, T. (1996). 3D geographic network displays. *ACM Sigmod Record*, *25*, 50–54. doi: 10.1145/245882.245901

Cui, W., Zhou, H., Qu, H., Wong, P. C., & Li, X. (2008). Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*, 1277–1284. doi: 10.1109/TVCG.2008.135

Dodge, M., & Kitchin, R. (2004). Charting movement: mapping internet infrastructures, *Moving People, Goods and Information in the 21st Century: The Cutting Edge Infrastructures of Networked Cities*, Routledge, pp. 159–185.

Dodge, M., & Kitchin, R. (2007). *Atlas of cyberspace*. Download from www.kitchin.org/atlas/contents.html

Guo, D. (2009). Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, *15*, 1041–1048. doi: 10.1109/TVCG.2009.143

Holten, D., & Van Wijk, J. J. (2009). Force-directed edge bundling for graph visualization. *Computer Graphics Forum*, *28*, 983–990, URL http://dblp.uni-trier.de/db/journals/cgf/cgf28.html#HoltenW09

Minard, C. J. (1864). Carte figurative et approximative des quantités de vin français exportés par mer en 1864. *Lithograph* (835 × 547).

Munzner, T., Hoffman, E., Claffy, K., & Fenner, B. (1996). Visualizing the global topology of the MBone. In *Proc. IEEE Symposium on Information Visualization (INFOVIS '96)*, IEEE, USA, pp. 85–92. URL http://dblp.uni-trier.de/db/conf/infovis/infovis1996.html#MunznerHcF96

Phan, D., Xiao, L., Yeh, R., Hanrahan, P., & Winograd, T. (2005). Flow map layout. *IEEE Symposium on Information Visualization (INFOVIS'05)*, 219–224, URL http://dx.doi.org/10.1109/INFOVIS.2005.13

Tobler, W. R. (1987). Experiments in migration mapping by computer. *Cartography and Geographic Information Science*, *14*, 155–163. doi: 10.1559/152304087783875273. URL http://www.ingentaconnect.com/content/cagis/cagis/1987/00000014/00000002/art00005

Tobler, W. R. (2003). *Movement mapping, center for spatially integrated social science*. URL http://www.csiss.org/clearinghouse/FlowMapper/