Movers: Exploring Refugee Movement with Social Media Data

Y. Xin[1], M. B. Simpson[1], Y. Xu [2], F. Sun [2], F. Linder[3], X.Liu[1]

[1]The Department of Geography, The Pennsylvania State University, University Park, PA16802, USA
Email: { ykx5030; marksimpson; xiliu}@psu.edu

[2]College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA16802, USA
Email: {yux115; fzs122}@psu.edu

[3]Department of Political Science, The Pennsylvania State University, University Park, PA16802, USA
Email: fridolin.linder@gmail.com

1. Introduction

Knowledge of refugee movements can help humanitarian response organizations better target and provide services to refugees, and help governments make informed decisions when planning housing, food, transportation, and other infrastructure provisions. However, understanding refugee movement patterns is difficult due to the transient nature of refugees, which makes conventional data collection methods difficult to implement, i.e., field surveys. Visual analytic tools can provide a novel perspective into refugee crises worldwide by enabling interaction with new data resources, such as social media.

Geo-referenced social media may provide important information as to the nature and rate of refugee movements. However, it is well known that social media data is "noisy"—that the amount of information is great, but the amount of relevant information may be very low. Accordingly, we designed an exploratory visual analytics tool, Movers, to investigate the feasibility of using geo-referenced social media for understanding refugee movement. This system is designed to support analysts in systematically exploring geo-referenced twitter data in order to identify refugees. The identification of individual Syrian refugees who fled to Germany is used as case study to test the functionality of this tool. Other European countries will be examined later for further analysis.

2. Refugee Movement

2.1 Syrian Refugee Crisis

According to the 1951 Refugee Convention (UNHCR 2001), a refugee is defined as someone who "owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality, and is unable to, or owing to such fear, is unwilling to avail himself of the protection of that country." Since 2007, the refugee population also includes people in a refugee-like situation. In our case study of Syrian refugees, refugees are defined as individuals who are fleeing to other countries and have not yet returned home.

The reason for urgent development of refugee movement analysis corresponds to the escalating refugee crisis in Middle East and Europe. By the end of 2014, 59.5 million people have been forcibly displaced. One in five are Syrians. Since 2011, there are more than four million Syrians fleeing to neighboring countries and Europe, making countries like Lebanon hosting 232 refugees per 1000 inhabitants (UNHCR 2015).

One of the biggest obstacles in making sense of refugee movement patterns is to employ macro and micro level data for service providers to act upon. Currently, the most used publicly accessible refugee movement data is the annual reports from United Nations High Commissioner for Refugees (UNHCR). The UNHCR disseminates yearly and monthly aggregate statistics of refugees by countries. However, understanding refugee movement patterns is difficult due to the transient nature of refugees and travel (often without official travel documents) through multiple countries via whatever modes of transportation are available.

Boyandin et al. (2010, 2011) have been exploring ways to visualize this dataset through origin and destination methods. The visualizations provide an overview of forced migration since 1975, the year the dataset was first available. Nevertheless, the lack of systematic detailed data of often highly dynamic refugee crises poses numerous challenges for hosting countries to accommodate influxes of refugees fleeing from persecution and conflict.

2.2 Movement Visualization and Modeling

Even though there are limited visual analytics studies targeting this area, related movement analysis in disaster response and migrations are available. Researchers have been exploring the basic laws of human movement patterns and found they are highly predictable (Gonzalez et al. 2008, Song 2010). Blumenstock inferred patterns of internal migration from mobile phone call detail records (CDR) in Rwanda (Blumenstock 2012). Despite the popularity of mining mobile phone data, it is impractical to use mobile phone data in understanding refugee movement patterns because of its country-specific nature and the lack of accessibility. However, the most popular alternative is publicly available social media data. Hawelka et al. explored geo-referenced Twitter data as proxy for global mobility patterns (Hawelka 2014). Blanford et al.(2015) studied cross-border movement with Twitter data. They were able to capture movement of people at different temporal and spatial scales.

Human mobility after an emergency may also be significantly more predictable than previously thought. Studies assumed that the mobility process to be either Markovian or non-Markovian (Poletto 2013). Song et al.(2014) showed that movement after large-scale disaster correlate with the movement pattern during normal times and are highly impacted by a person's social relationships, the intensity of disaster, the damage level, availability of government appointed shelters, news reporting and etc. This discovery provides us directions on designing the system in order to reveal characteristics of refugee movements.

3. The Movers Tool

3.1 Overview

The system is implemented as a web-based application running in a browser. The interface consists of three main parts, feature selection, tweets information panel and visualization, and is described in more detail in section 3.4. Functionality of our system is designed to balance the information complexity and perceptual needs by including different visual panels. The raw tweet data are collected through a web crawler, which is then preprocessed before being read by the system, and our specific methods are detailed in section 3.2. In the case study illustrated in section 4, we focused specifically on Syrian refugees who fled to Germany during 2015, but the interface can be used on any appropriately formatted set of geo-located tweets. While our ultimate goal is to allow users to identify and predict forced migration patterns, our initial implementation focused on establishing the feasibility of using social media such as twitter to find refugees' movement. This entails finding individual users who appeared to be refugees, by removing users who were not, such as journalists or tourists, who also travel internationally, and examining the content of the individual user's tweets.

3.2 Data Processing

The raw twitter data were collected with crawlers for the whole year of 2015 and cover one-quarter of the whole world. We preprocessed about 3 TB raw data first by selecting users who has at least one geo-located tweet in the area of Syria and the area of Germany (defined by bounding boxes). This step reduced the data to approximately 1.5 million tweets generated by 1560 users. We then created a file with features of users and tweets in JSON (JavaScript Object Notation) format, which is about 150 MB in size. The tweets JSON file contains coordinates of each tweet, country, language, speed information and so on. User JSON file organized by each user with a sequence of coordinates of places visited. This file can be directly loaded into client memory and further processed through web browser. The contents of tweets are stored separately on a server, which supports querying detailed information about specific users and their tweet contents through a MongoDB database.

3.3 Design Rationale

Our first goal was to identify Syrian refugees via geo-located twitter data. We treat this as a visual data mining operation, where the user is seeking to identify a very weak "signal" in an extremely "noisy" dataset, full of potential false positives. We follow Shneiderman's information seeking mantra (1996) in the design of the interface, by providing an overview of the data via the map and "time travel" views, a suite of filters, the ability to select user trajectories to view individual tweets and display of selected users' tweets in the information panel. See Figure 1 for the workflow of our system. We place a strong emphasis on filtering through the side panel, which in addition to allowing users access to the filters themselves, keeps track of how many users they have excluded so far. We mainly provide for subtractive filtering rather than additive filtering to assist users reduce irrelevant information.

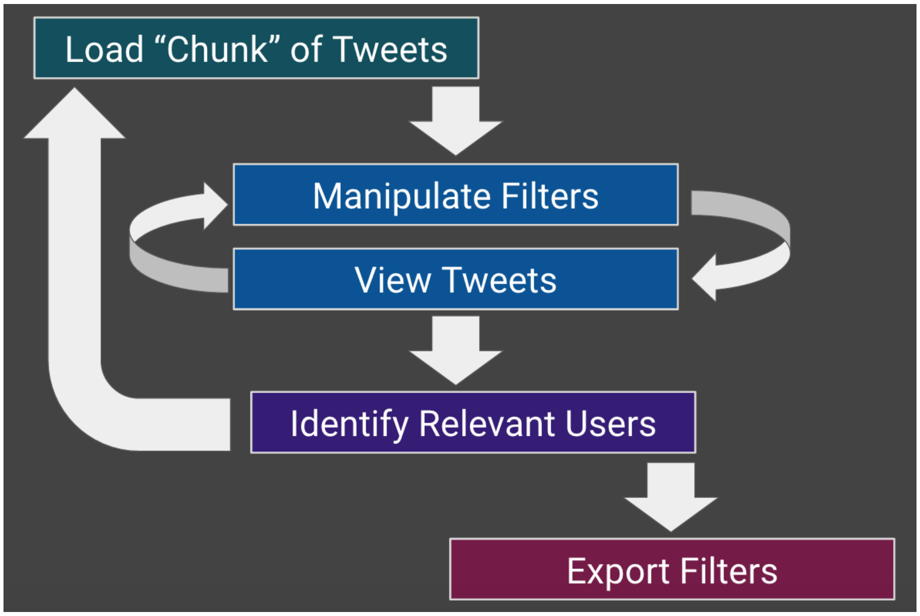

Figure 1. Workflow of the system. Data are added into the system in chucks (by storing order in the database) to reduce processing time and visual clutter. Users then manipulate filters to remove superfluous users (i.e., non-refugees), using the text of the tweets to verify user types. Once a set of users of interest (UOI) is identified, those filter settings can be applied to the next set of data, or saved for future use.

3.4 The Movers Interface

The user opens the interface in a web browser. There are currently three main sections: the feature selection (filter) panel on the left, the tweet information panel in the middle showing tweet contents of selected twitter users, and visualization panel on the right, with map view (Figure 2) on the top and time travel view (Figure 3) at bottom navigated by a scrollbar.

The feature selection panel contains the filter controls. The following filters are currently implemented: maximum user speed, number of countries visited by a user, languages (to exclude), countries (to exclude), and countries that the user must have passed through. ("include"). The top displays the number of users shown versus those removed by the filter, as well as the total number of users. The bottom is reserved for importing and exporting filter settings. This gives users the option of using configured filters to identify similar potential refugees.

The map view shows the trajectories of users, with their tweets forming vertices of the trajectories. The trajectory colors are used to display additional user attributes, such as twitter profile language and travel speed. A GeoJSON-Vector-Tile JavaScript library is used to slice geojson data into vector tiles on the fly and display large amounts of data smoothly and seamlessly. Simplification of trajectories is automatically calculated based on the zoom level. Leaflet JavaScript library is used to provide basemap functionality. Currently all individual trajectories are shown, but visual aggregation has been tested and will be implemented soon.

The time travel view focuses on time rather than space, since the temporal signature of refugees can be very different from other users (they may walk instead of fly, or spend weeks waiting to cross a border). User trajectories in the time travel view are constrained on countries showing in the y-axis with consistent color as trajectories in the map view. In default setting, Syria is placed close to bottom and Germany at the top. Users can reorder the countries along y-axis to explore movement patterns. Once a refugee is identified, its country visiting pattern can be applied to reorder the countries, to assist finding of other potential refugees.

The x-axis is reserved for time, so that users can be visualized as moving from left to right. In this view, steep slopes indicate faster travel speeds, and horizontal lines indicate no travel outside the country. Refugees travelling overland to Germany therefore appear as a gentle sloping line interspersed with horizontal "stays" in various countries. Two JavaScript libraries are used for this visualization: D3 enables the system to generate interactive visualization and dynamic filtering, while Crossfilter speeds up the process of exploring large multivariate data.

Both views enable user interactions with trajectories. Left click to show all tweets for the selected user and right click to remove user. The linking between two views is currently under development, which is designed to respond to user's mouse over simultaneously.

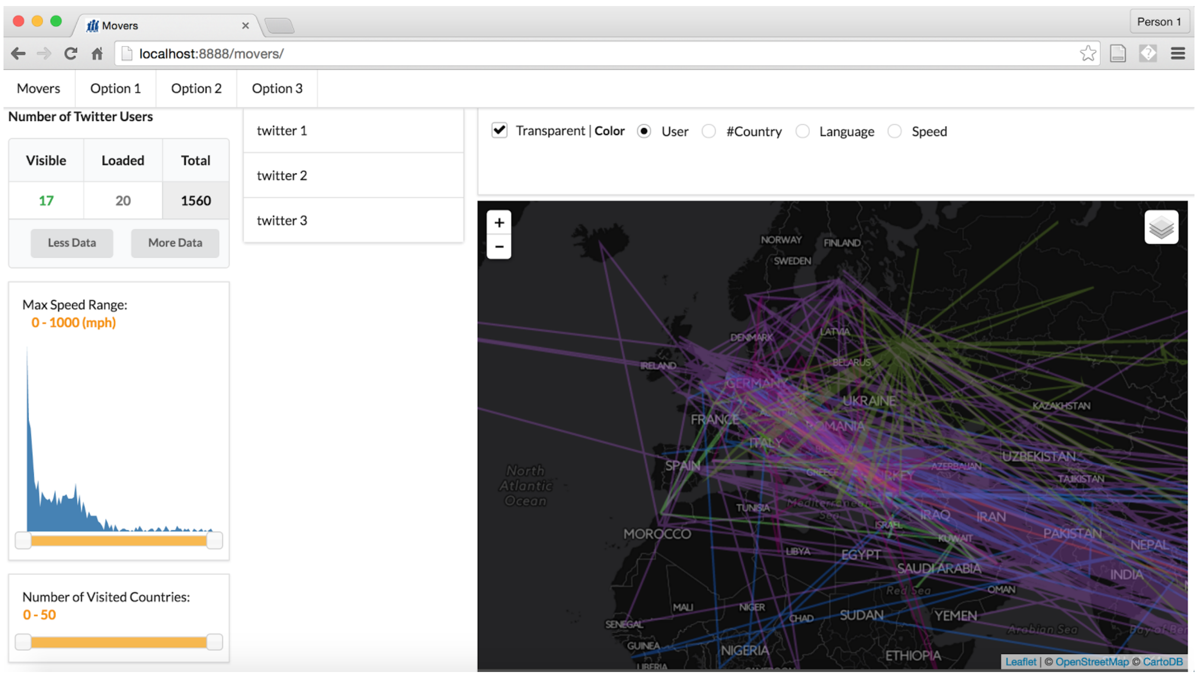

Figure 2. Main page with map view. a) Left side is feature selection panel. From top to bottom, the filters are Number of Twitter Users, Maximum Speed Range, Number of Visited Countries, Exclude Language, Exclude Countries and Include Countries. b) Middle part is Tweet Info panel, showing tweet contents of selected users. (In progress) c) Right side is Visualization panel showing map view. Four basesmaps are provided, including grey scale, dark matter, satellite and street map.

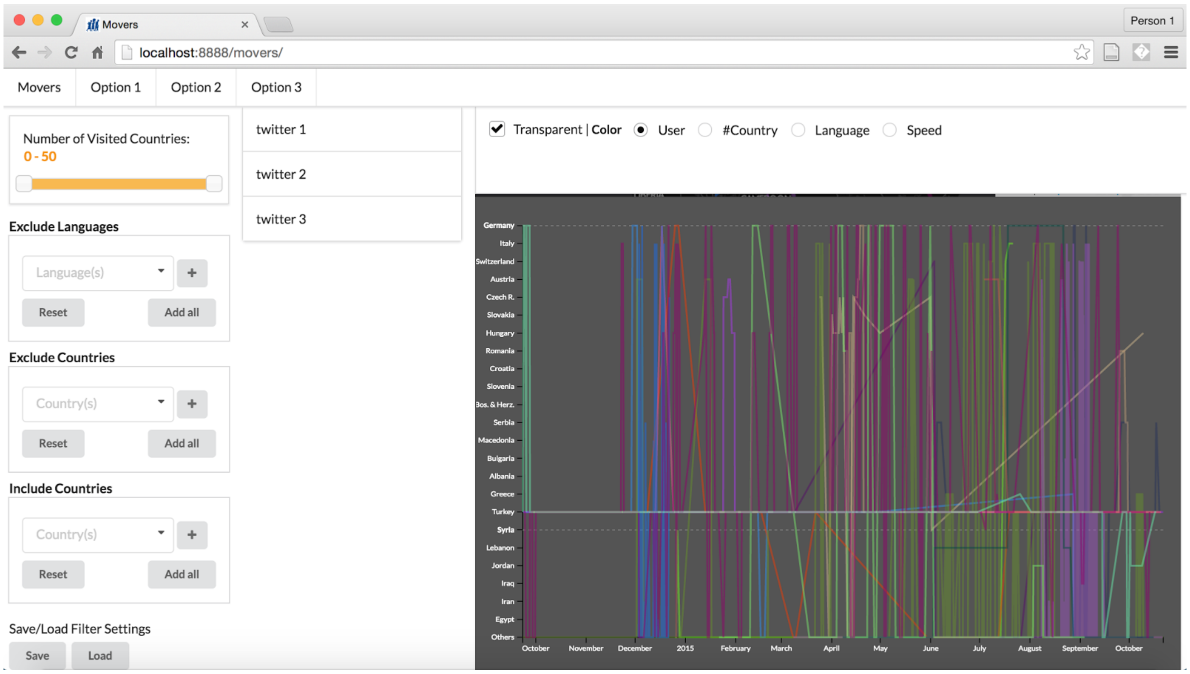

Figure 3. Main page with time travel view. a) b) Same in Figure 2. c) Visualization panel showing Time Travel view. X-axis is time, Y-axis is country visited with the middle dash line representing Syria.

4. Case Study

4.1 Scenario

As refugees defined as individuals fled outside of their home country for the fear of persecution and violence (UNHCR 2015), we anticipate their fleeing trajectory as moving towards a safe country where threats are not present. When focusing on the mobility of Syrian refugees towards Germany, we assume that they move slowly across borders from Syria to Germany with Arabic as the dominant language used.

So we apply the following filters to find potential Syrian refugees through Twitter data.

- speed range: 0 - 300 miles per hour
- number of countries visited: 0-20
- excluded country: other countries
- included country: Syria

After applying these filters, similar trajectories are shown in both map view and time travel view. After reviewing the content of their tweets, we successfully identified refugees who said "I am a Syria refugee" in one of his/her tweets.

5. Discussion and Future Work

As a visual analytic tool created for users to identify refugee movement patterns at a micro daily level, our case study result showed the effectiveness of our tool. We are able to identify refugees and their fleeing patterns.

This tool is an exploratory tool that is used to explore what a great design should be like rather than designing a perfect tool with one attempt that is able to achieve all the intended purposes. As such, even though we evaluate the effectiveness of Movers through a case study, we put more emphasis on the iterative design and development processes. Features are under development or to be developed at next stage:

- Fusing heterogeneous data, from news, UNHCR, or other corpus

  To make the best of both macro and micro analyses, integration between the two will be implemented. Showing statistical numbers of refugees as background reference not only assists the identification of potential refugees, but also serves as evaluation of our tool.
- Expand time period to have comprehensive look at forced migration patterns

  The Syrian crisis started in spring 2011. Over the past five years after the war began, almost 5 million Syrians have registered or are awaiting registration with the UNHCR. In current study, the data we used only covers 2015. By expanding the timeline, we expect to find more refugees through the data exploration.
- Interactive learning with machine learning techniques

  The initial identification of refugees relies mainly on users' judgements via a set of filters and tweets analysis. By adding machine learning techniques, after users identified certain refugees, they could use identified movement pattern as training set to find other potential refugees with similar movement patterns.
- Expand to identify other movement patterns, such as business or international journalists.

  This tool can be potentially used in other applications involving visual analytics of movements.

Acknowledgements

We would like to thank Alan MacEachren and Alexander "Sasha" Savelyev for their invaluable guidance, and Scott Pezanowski and the GeoVISTA Center for their help in acquiring data.

References

Blanford J. I., Huang Z., Savelyev A., and MacEachren A. M., 2015. Geolocated tweets. Enhancing mobility maps and capturing cross-border movement. *PLoS ONE*.

Blumenstock J. E., 2012. Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development*, 18(2):107–125.

Boyandin I., Bertini E., and Lalanne D., 2010. Visualizing the World's Refugee Data with JFlowMap. *In Eurographics/IEEE-VGTC Symposium on Visualization*.

Boyandin I., Bertini E., Bak P., and Lalanne D., 2011. Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data. *In Eurographics/IEEE-VGTC Symposium on Visualization*, volume 30, pages 971–980, 2011.

Gonzalez M. C., Hidalgo C. A., and Barabasi A.-L., 2008, Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hawelka B., Sitko I., Beinat E., Sobolevsky S., Kazakopoulos P., and Ratti C.. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41(3):260–271, 2014.

Poletto C., Tizzoni M., and Colizza V. 2013. Human mobility and time spent at destination: Impact on spatial epidemic spreading. *Journal of Theoretical Biology*, 338:41 – 58.

Shneiderman, B. The eyes have it: a task by data type taxonomy for information visualizations. *In 1996 IEEE Symposium on Visual Languages*, 336–43.

Song C., Qu Z., Blumm N., and Barab´asi A.-L.,2010. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

Song, X.; Zhang, Q.; Sekimoto, Y.; and Shibasaki, R. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 5–14. ACM