Name Profiler Toolkit

Feng Wang, Brett Hansen, Ryan Simmons, and Ross Maciejewski ■ Arizona State University

Name data provides an easy entry mechanism to relate population-specific information to aggregate demographic data. Census data, phonebooks, social networking profiles, and other sources provide full names in readily available data formats. In a similar fashion, demographic information can be collected from census data, surveys, and similar sources. Such data has been widely explored as tools for marketing and advertisement, while other uses of name data have historically played a more lighthearted role—for example, what was the most popular name in the year you were born? Given that family names (surnames) and forenames are widely used markers for spatially referenced population data, tools that can explore and link these names to other data sources can provide analysts with novel means to explore and hypothesize about the geography of names. For example, Pablo Mateos and his colleagues revealed cultural naming practices for new and existing communities,1 and James Cheshire and Paul Longley developed methods for classifying surname distributions in Great Britain.2 Such work has led to public tools such as the worldnames.publicprofiler.org. (See the “Related Work in Name Data” sidebar for more details.)

In recent work, we augmented such name exploration work with links to secondary data sources, including income distributions from the census bureau.3 In this article, we further develop that methodology to analyze spatial probability distribution functions and explore visualizations to analyze linked secondary data sources, for example, linking forenames to estimated housing prices. Such work can extend to other categorical spatial temporal data (such as crime and health). Joint probability measures between different categories can provide analysts with insight into where categories overlap, and secondary data sources could provide insight into the “why” behind that overlap.

To demonstrate our contributions, we extended our previous version of the Name Profiler Toolkit4 to enable cross-data exploration and joint probability distribution. This work illustrates how data science methods can help enable the exploration of geographically distributed phenomenon and provides insight into how multisource data can reveal secondary characteristics. By linking name, income, and age, analysts can explore unique geographic profiles of the United States to reveal relationships between locations, ethnicity, and income divides. The Name Profiler Toolkit is available online at vader.lab.asu.edu/NameProfiler.

Task Analysis and Design

With point process data (such as telephone records and criminal incident reports), one of the main questions being asked is “How are the data spatially distributed?” However, once various patterns are identified, more complex questions might arise such as “What else in the data has a similar distribution” or “What features are driving these patterns?” In this article, we focus specifically on surname and forename distributions and on tasks associated with exploring such data, but all tasks and methods we discuss could be applied to other point process data.

We begin our discussion by identifying several critical analysis tasks and discussing system requirements to support them:

- **T1: Interactive distribution exploration.** Given categorical spatial data (such as name-address pairs...
Surnames are commonly used to explore the dispersion and heterogeneity of populations around the world, and researchers have created a variety of interactive maps to visualize global surname distributions. For example, Yifan Zhang and his colleagues utilized a database from the WorldNames Project to link IP addresses, locations, and surnames to visualize possible immigration routes across the globe. James Cheshire and Paul Longley described the spatial concentration of surnames within the United Kingdom. Their work identified core surname areas for 92 percent of the surnames analyzed. The other 8 percent of surnames were dispersed and hard to pinpoint within centralized areas. However, even with weighting to diminish the influence of urban centers, their procedure revealed high-density regions of common ethnic surnames within these urban areas. They hypothesized that this arose from migration patterns and that it indicates homogenization within regions.

Pablo Mateos and his colleagues explored personal naming networks, looking for name clusters based on ethno-cultural customs and social norms. Their research built naming networks utilizing existing population registers, including an Auckland, New Zealand database for preliminary testing and the WorldNames database for worldwide results. In preliminary testing, the research created a complete naming network of Auckland. The researchers found clear clusters for particular names and that naming clusters preserve through migration. Previous research has also analyzed the social mobility rates of surnames within the United States. Gregory Clark investigated the social mobility between generations in the United States from 1920 to 1949, 1950 to 1979, and 1980 and 2012. Clark categorized surnames into different classes and found that social mobility between 1920 and 2012 was much less than expected. Surnames can also be used to describe class structures within regions, for example, linking a person’s ethnicity to the reported income values in a region. Previous census research has shown a correlation between ethnicity and income. Carmen DeNavas-Walt and her colleagues found people of black or Hispanic origin made significantly less than those of white or Asian descent. Their work also noted that the income disparity persisted between the ethnic groups, with white and Asian people earning more yearly income than their black or Hispanic counterparts. This research also indicated that Americans in the western and northeastern portions of the United States had the highest median income, followed by the midwest. The south had the lowest median income of the four regions of the United States. Denavas-Walt’s findings support the theory that income differences exist between geographic regions and ethnicities within the United States. Thus, the ability to link names (which often contain ethnographic information), age, and income can provide powerful insights into geographic structures.

References


in the United States), a common question is “What is the spatial distribution of name x?” To address this aggregation requirement, the system must interactively retrieve the probability distribution for a given category (in this case, a name) and interactively return a rendering of the probability distribution of any user-selected category.  

**T2: Distribution similarity exploration.** Once a given distribution is obtained, the next obvious question is “What other categories (such as names or types of crime) have a similar spatial distribution?” Although the example dataset here uses names, any categorical point process dataset could have similar questions. For example, “What crimes have a similar spatial distribution to this crime?” or “What diseases have a similar distribution?” Thus, the system must provide insight into related data categories and let the user navigate through these complex relationships.

**T3: Joint distribution exploration.** Although identifying similar distributions is critical, questions about where these distributions have the highest overlap are also of interest. For example, “Where do traffic accidents and drunk-driving arrests commonly co-occur?” or “Are there unique Irish-French-Italian immigrant settlements?” Therefore, the system should support joint probability distribution analysis.

**T4: Linked data-source exploration.** Upon exploring such point process data, another analytical task would be to link these data with other related data sources, such as income data. This
would allow people to explore, for example, the income profile of surnames and forenames, age attributes, or other spatially related phenomenon. To address this, the system must support methods for visualizing secondary data sources.

This article focuses on methodologies for linking spatial distributions (such as names) to secondary data sources (such as income) to enable advanced analyses. Furthermore, we focus on joint distribution exploration to enable more complex analysis questions. Finally, because of the relatively large size of the data, we also develop methods for speeding spatial queries for similarity analysis.

Name Profiler Toolkit
The Name Profiler Toolkit is a geo-visual analytics system for exploring name distributions. We utilize publicly accessible telephone data and link this data to US Census and Zillow data. Building off of previous research questions and designs, our goal is to enable both novices and experts to explore name distributions and spatial relationships. We focus on three issues: multisource data fusion, similarity metrics, and joint distribution analysis.

The system interface consists of three primary views: the name-density geographic view, the similarity wordle, and the name-name heatmap view. Figure 1 provides an overview of the system, showing the name-density and wordle views. Here, the user is exploring the surname Rossi and can immediately identify the strong east coast pocket of Italian immigration centered in New York City. The wordle view is displaying names with similar spatial distributions, and although many Italian names are prevalent (such as Borrelli and Marchione), we see this distribution covers a variety of ethnic surnames, perhaps giving insight into how immigrants migrated after coming to Ellis Island.

Multisource Name-Age-Income Data
In this work, we use the 2008 US public telephone directories as the primary data source. These directories contain more than 78 million records with forename/surname pairs and their associated street addresses. These addresses are geocoded using the Google API and can be further aggregated to geographical units, such as census block, census tracts, and zip codes. To enhance the telephone directory data, we also captured age information on forenames. The ages are estimated from the US Social Security Administration (SSA) baby name database and actuarial tables. The actuarial tables estimate how many people born in a given year are still alive. Overall, our database consists of 1,426,633 unique forenames and 1,649,469 unique surnames.

We also collected data related to household income. The income distributions were estimated from the household income in the 2008–2012 American Community Survey Five-Year Estimates. This dataset describes the ratios of different income buckets within each census tract. To further refine the income data, we estimated the real estate value distributions from the data provided by Zillow.com. We used the 2016 real estate trend data published by Zillow Research (www.zillow.com/research/data/) as well as estimated housing
prices collected from the Zillow search API. This data fusion partially supports T4 by building the necessary data tables.

**Name Density Estimation**

To visualize the spatial distributions of different names, we employ fixed-bandwidth kernel density estimation (KDE). The multivariate KDE can be defined as

\[
\hat{f}_h(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h^d} K\left(\frac{x - x_i}{h}\right),
\]

where \( h \) is the bandwidth, \( N \) is the number of samples, and \( d \) is the number of dimensions. In our application, we use a symmetric, 2D Epanechnikov kernel for fast computation. The Epanechnikov kernel is defined as

\[
K(u) = \frac{3}{4\pi} (1 - u^2)1_{[u<1]},
\]

where \( 1_{[u<1]} \) evaluates to 1 if the inequality is true and 0 otherwise.

However, the distributions of most names are highly correlated to the distribution of the US population. To measure the relative spatial probability, we apply Bithell’s risk function as a means of normalizing the name density distribution with respect to the overall population. In this manner, we can identify local hotspots. Bithell’s risk function is defined as

\[
p(x) = \frac{\hat{f}_i(x)}{g(x)},
\]

where \( \hat{f}_i(x) \) is the density of distribution \( i \) and \( g(x) \) is the density function of the entire US population. The left side of Figure 1 shows the probability density estimation of the surname Rossi. We use KDE primarily because of its ability to parallelize the algorithm to enable near-interactive visualization rates in order to support T1.

**Joint Distribution Analysis**

Although KDE supports the exploration of spatial distributions, further extensions are necessary to identify co-locations between names. For example, if we know the distribution of the forenames and surnames, can we estimate the distribution of people with this full name? If you know your mother and father’s surnames, can we see where in the United States they were most likely to meet? Our goal is to visualize the joint probability distribution between multiple names to enable a variety of new analysis questions. To visualize these overlaps, we use the Bhattacharyya coefficient to measure the similarities between the distributions of names. For two distributions \( p \) and \( q \), the Bhattacharyya coefficient is defined as

\[
BC(p,q) = \int \sqrt{p(x)q(x)}dx.
\]

This coefficient can be generalized to more distributions using the following extension:

\[
BC(p_1,\ldots,p_n) = \int \prod q_i(x)dx.
\]

Figure 2 shows the joint spatial distribution of Dubois, Murphy, and Lund, which are some of the most popular French, Irish, and Scandinavian surnames in the US, respectively. In the Dubois distribution map, we can see the heatmap centers strongly on Boston as well as a hotspot near New Orleans (a well-known French settlement). For Murphy, we can also identify a Boston hotspot, but it extends down to New York City and has a strong presence in Chicago. Finally, when exploring Lund, we can see the Minnesota immigration wave and a large hotspot in Wyoming.

Finally, if analyst wants to explore where French, Irish, and Scandinavians are most likely to co-locate in the US, the Bhattacharyya coefficient can be used to provide a straightforward metric to visualize the spatial probability of overlap. Thus, Figure 2d shows that the most likely regions of overlap include Boston, northern Maryland, Chicago, and Minneapolis. Even though New Orleans is a hotspot for Dubois, we can see that it is completely removed in the joint distribution, and the impact of New York City is also greatly lessened. Thus, such a methodology directly supports T3.

However, the Bhattacharyya coefficient calculation has a high computational complexity for nearest-neighbor searching and clustering. The Bhattacharyya coefficient combined with the name-density estimate view can provide insight into the total joint spatial relationships of names. We have also developed a methodology to investigate the sublevel name-name relationships (either forename or surname pairs) using a pixel-based heatmap view. Each name can be represented as a vector describing either the spatial distribution, income distribution, or age distribution (forename only). Using these vectors, the Name Profiler Toolkit calculates the 400 nearest neighbors for each name. The similarities between these sets of neighbors are measured with a Jaccard index:

\[
J(A,B) = \frac{|A \cap B|}{|A \cup B|}.
\]
where A and B are two sets of nearest neighbors, $J \in [0, 1]$. Figure 3 shows the Jaccard index heatmap for six different common Asian surnames using the spatial distribution as the underlying similarity vector. The darker the pixel is shaded, the more similar the name-name relationship. Here we can see that the surnames that originate from the same countries (such as the Vietnamese surnames of Tran and Nguyen and the Chinese surnames of Chen and Wong) have a stronger spatial relationship than when compared with those originating from another country. This view was designed to further support T2, T3, and T4.

**Linked Secondary Data**

While spatial estimates give us insight into locations, we might also wish to analyze secondary information via data fusion. In this section, we illustrate the application of data fusion for enhancing the name-distribution analysis.

**Name-age estimation.** One question an analyst might have when exploring name data is “What are the age relationships between names?” To estimate the distributions of the ages (forenames only), we adapted the methodology proposed by Nate Silver and Allison McCann. The SSA baby name database provides the number of newborn babies along with the frequency of a given forename. The SSA also provides life tables that contain the number of people born in a given year who are still alive. These data are provided in decade-based sets. The years between decades are interpolated between the neighboring decades. The age distributions can be calculated with the overlaps between these two datasets. Silver and McCann predicted a name’s median and quantiles to compare names. However, their method ignores the distributions of the names. For example, the bimodal distribution of the forename Violet would not be observed with the median and quantiles. Thus, we expanded on their methodology by splitting the age range into 10 buckets from ages zero to 100 (older than 100 is grouped into the 100 bin) to create a histogram of the forename distribution. After normalization, the histogram vectors are indexed for similarity search.
The wordle view is then extended to allow users to interactively modify the color or size of the words to represent the similarity between name-location, name-income, and name-age (forename only) relationships, thus supporting T2 and T4.

**Income estimation.** Along with name ages, the income distributions of surnames and forenames can also be estimated from census data. To accomplish this, we mapped the telephone directory records to census tracts. The census tract income data provides a 10 income segment histogram. To link the name data to the income data, we use a linear system analysis method. The linear system can be defined as $DX = B$, where the matrix $D$ is the histogram matrix of geographical name distributions. $D_{ij}$ is the number of name records for the $i$th census tract and $j$th name. The matrix $B$ is the histogram matrix of geographical income distribution. $B$ can be defined as $B = [B_1, ..., B_n]^T$, where $B_i$ is the 10-dimension vector of the income histogram of census $i$, which is from the census dataset.

Because $D$ is guaranteed to be full-rank, we use a nonnegative least-square solver to solve the system. Finally, the distribution vectors for each name are normalized for similarity indexing. Figure 4 shows that the richest surnames (by census block income) are most likely to be in New York City. From our pixel-based heatmap view, we can also identify three Italian names (Ellantoni, Socci, and Bueti) that have similar income distributions. The top four names and the sixth name, Doniger, show a cluster in terms of the overlap between income and spatial distributions.

Although using the census block to estimate income is reasonable, our previous work on name income estimation found that using the census block measures to estimate income often proved to be too coarse of a representation. Therefore, we explored linking estimated the real-estate values for addresses. We estimated the home value of

Figure 3. Jaccard index heatmap for the top six Asian names in the United States. Nguyen and Tran originate from Vietnam. Lee has origins in China, Korea, and Europe. Kim originates from Korea. Chen and Wong originate from China. The nearest neighbors are estimated from the spatial distribution images. A darker red pixel indicates that the names are more similarly distributed across the United States.

Figure 4. Spatial distribution of the richest surnames: (a) joint spatial distribution and (b) heatmap showing the overlaps between neighbors in terms of spatial distribution. Here, we see the places where one is most likely to meet a high income surname is New York City and Miami, but we can also see that the spatial distributions of these names are not highly correlated, with only Socci, Bueti, and Bellantoni sharing any high degree of spatial similarity.
each telephone directory record using Zillow home value index (ZHVI) data and estimated values collected through the Zillow search API. However, because of the rate limit of the Zillow API, only a portion of the street addresses can be queried. For each census block, we query at least one address using the API. For each record in the database, the real estate value is then estimated as follows:

- If a valid result is returned from the search API, the Zillow estimated value is used as the home value.
- If there are results from search API for other records in the same census block, the median value of the estimates is used as the home value.
- If no results are returned from the search API, the ZHVI data is used.

We found that the estimated income values follow a skewed distribution. To normalize the data, we assigned the estimated values into 10 quantiles (deciles) and then calculated the distribution vectors for each name. Surprisingly, we found that the spatial distribution of the names with the highest housing values disagrees with the distribution of the richest names in Figure 5. Obviously, housing prices and salaries are regionally dependent, and we do not adjust for cost-of-living estimates or other factors. Such adjustments should be included in future work for a more complete analysis.

**Similarity Indexing**

Although a major focus of the Name Profiler Tool-kit is combining multisource data, another major design factor is interactivity (T1). When we compare the spatial distributions of names, our system is actually comparing the visualized KDE images. We utilize image search methods to improve our similarity searches to enable near real-time interactivity.

In our system, the KDE for the spatial distribution is approximated by first splitting our map into bins equal to the image space. For example, if we have a 256 × 256 map, then the names are aggregated into 256 × 256 bins. This means that the computational cost is directly correlated to the resolution of the result image. Our observations indicated that low-resolution images (55 × 94) provide a reasonable estimate of the general distribution patterns of the name records. Thus, we use a low zoom level to generate feature vectors for nearest-neighbor search. For visualization, however, the low-resolution images lack detail when zooming. Therefore, we also employ a multilevel spatial hashing index over the records to accelerate the spatial KDE calculation at different zoom levels\(^{14}\) for higher-resolution visualization.

For similarity analysis, the lowest zoom level KDE of the names is calculated. The similarities between these distributions are defined as the similarities of these KDE hotspot images. In our previous work, we measured this distance with L2 distance over the image vectors.\(^{3}\) This method suffers from high computational complexity. Furthermore, the curse of high dimensionality also impacts the accuracy and indexing performance.

To address this, the updated Name Profiler Tool-kit only stores and indexes the most representative image features. Specifically, we decompose the KDE images into linear combinations of several

---

**Figure 5. Distribution of the top five names with highest housing values from Zillow:** (a) joint spatial distribution and (b) heatmap showing the name-name relationship with respect to their underlying spatial distribution. Unlike the census income map, here we see the home income concentrated in Los Angeles, and we see a relatively strong spatial similarity among most of the names, except for Petrossian and Petrosonian.
eigen images. Because the dataset is static, we do not need to store the eigen images, and only the proportions are used in the indexing. To compress the proportion vectors, we perform principle component analysis (PCA) for dimension reduction. Because of the large size of our data, we use an incremental PCA implementation to reduce the space complexity. With the compressed feature vectors, we can use the same indexing schema with other feature vectors for nearest-neighbor search. The distribution vectors are indexed using a fast-approximate nearest-neighbor search (FLANN).

Previously, the precalculation of KDE image vectors and similarities was a bottleneck between the big name database and the limited number of names we were able to use. Now we can calculate the incremental PCA results for all names (both surname and forename) with more than 100 records within 4 hours. The transformed vectors are used as feature vectors for indexing. We tested our system on a Linux server with a four-core Intel i7 2.67 GHZ CPU and 20 Gbytes of memory. The average response time for nearest-neighbor queries was 11 milliseconds, and the average response time for a KDE image-generation query was 3 seconds. Most of the runtime memory was consumed by the nearest-neighbor query method. The system took 2.4 Gbytes of memory for the indexing data structures. To obtain higher performance, we cached the most popular queries. Such a methodology could be employed for any categorical spatial (or spatiotemporal) data.

Case Studies

To demonstrate the Name Profiler Toolkit’s flexibility, we present a series of interesting name distributions from across the United States, and we encourage readers to utilize the tool at vader.lab.asu.edu/NameProfiler and make their own discoveries as well. When demonstrating our system, we observed that users typically want to first explore their own names and observe the spatial behaviors of their surnames or look at the joint probability of their forename-surname pair.

Chinese Immigration Waves

In our first example, the primary author of this article (Feng Wang) wanted to explore common Chinese surnames. Depending on the time period during which families immigrated, their Chinese surnames have undergone different romanizations. The surnames Wang and Zhou in Hanyu Pinyin romanization are spelled Wong and Chow in the scheme used in Taiwan and Hong Kong. The Taiwan romanizations were also used by the first waves of Chinese immigrants to the United States in the 19th century. In Figure 6a, Wang and Zhou (the Taiwanese wave) are input into the system, and the joint distribution map is displayed. This is repeated in Figure 6b for the Hong Kong wave (Wong and Chow). The results show that there are no obvious differences between these two groups, with both groups distributing mostly to large metro areas such as San Francisco, Los Angeles, and New York.

Figure 7a shows that the Zillow housing value distribution can be used as a clear indicator between these two groups. These four names form two clusters according to their ethnic origins, and we can compare the wordle views between these groups. In Figure 7b, we can see that the similar names are also mostly Asian names romanized in the Hanyu Pinyin. In Figure 7c, we can find more names romanized with the Taiwan scheme. This implies that the two waves of immigrants have unique income profiles, even though they are spatially embedded in similar locations.
This analysis began with the interactive analysis of two surnames (T1) and proceeded to explore distribution similarity (T2) and joint probability distributions (T3). Finally, by linking secondary information (T4), the user can identify an interesting social phenomenon of income distribution differences based on different romanizations of the same Chinese surname.

### Income Analysis

While exploring one’s own surname was the primary use of the Name Profiler Toolkit system, our users often wanted to ask this question: “What is the richest or poorest name?” For any name, we can calculate the median income from our various metrics. To support this analysis, we designed the system to automatically load the top 10 richest and poorest names.

In this example, we compare the geographical distributions of the 10 surnames with the highest and lowest incomes across the United States. Figure 4 shows that the richest names (based on census income metrics) distribute jointly in New York City. Figure 8 shows the joint spatial distribution of the top 10 surnames with the lowest income. The joint distribution map indicates that these surnames are most likely to be found together near Atlanta, Georgia; Detroit, Michigan; and southern Mississippi.

Figure 7. Income difference between the two Chinese immigration waves: (a) clusters from the name-name housing value heatmap view, (b) an income distribution of surnames similar to Wang and Zhou, and (c) an income distribution of surnames similar to Wong and Chow. In the wordle, the color indicates the similarity of the housing price distribution, and the font size indicates the similarity of the census block income. The two immigration waves have unique income distribution profiles that seem likely to be dictated by the time period in which they immigrated.

Figure 8. Distribution of the top five names with lowest housing values from Zillow: (a) joint spatial distribution and (b) heatmap showing the name-name relationship with respect to their underlying spatial distribution. The most likely regions to jointly encounter these names are Atlanta, Georgia; Detroit, Michigan; and southern Mississippi.
names all indicate high amounts of similarity between nearly every one of the 10 name-name pairs. In the top 10 richest names, only a small subset of the names has a similar income profile.

This analysis followed a different task path. The first task relied on finding name distributions based on secondary data properties, but this feature is only supported by providing the top 10 names associated with a data feature. This indicates a missing task from the originally identified task list and the need for future expansions. Once names are identified, a joint distribution is created (T3) and explored (T1). Currently, only spatial distribution analysis is supported, but future extensions to the Name Profiler Toolkit will implement a sparkline view to show the histogram distribution of income and age profiles.

The design aspects of the tool are well known, but the combination of the Name Profiler Toolkit’s statistical methods and multisource data fusion creates a powerful means of analyzing name distributions across the United States. This toolkit serves as an example of how data-science methods can address the transition from data to knowledge by integrating and extracting information, aggregating data for querying and visualization, and coupling geographic and information visualization methods together to enable the analysis of name-age-income-location relationships across the United States. Various implementations of this system have been shared over the course of two years, and anecdotal evidence suggests that the data matches users’ mental models. System users typically explore the tool for 10 minutes or more.

Users have suggested that we add a feature that allows them to query by drawing a custom distribution. Currently, the wordle is generated from the neighbors of average vectors. However, this method might not represent the real distribution of the selected names if the names have dispersed vectors. Users have also suggested that we include a feature for exploring the similarity network between the names and the feature vector selection to be used in wordle generation. Another potential extension is to add more similarity measures between the names and dimensions.

Although the visualizations presented in this work are standard, the implementation of a web-enabled system for large-scale visual analytics is still challenging. Our design has evolved from our previous work of precomputing similarities for a large number of categories to utilizing image-processing modalities. However, the methods we describe here focus on static data, and the multisource aspects of the data are handled on the backend. Developing systems that can adapt to multisource suggestions from the end user and enable data linkage and exploration into a seamless workflow remains an open area of exploration in the visualization community. By engaging with an online community, various user studies can be designed to evaluate effective visualizations and begin developing design principles for joint probability distribution analysis and linked data sources.

Future work will include integrating user-interaction tracking to capture provenance information to support user studies. Specifically, we are interested in what people actually would do given such a system with limited instructions. Traditionally, user studies have focused on how quickly a user can locate something or answer an analytical query given visualization tool A versus visualization tool B. We have a relatively open-ended framework, however. Anecdotally, users begin the process by searching their own name. Other features (such as richest surnames) are the result of questions the users wanted to explore after that initial exploration. A combination of paired analysis and provenance analysis would provide an interesting starting point for exploring use patterns in such a visual analytics environment. If common use patterns were to emerge, this could help identify search strategies to be shared that may inform other domain questions.

References
12. N. Silver and A. McCann, “How to Tell Someone’s Age When All You Know Is Her Name,” FiveThirtyEight, 29 May 2014.

**Feng Wang** is a research scientist at General Electric. This research was performed while he was PhD student at Arizona State University. His research interests include data visualization, visual analytics, and network data analysis. Wang has a PhD in computer science from Arizona State University. Contact him at fwang49@asu.edu

**Brett Hansen** is an undergraduate student in Computer Science at Arizona State University. His research interests include data visualization and data mining. Contact him at bjhanse1@asu.edu

**Ryan Simmons** is an undergraduate student in computer science at Arizona State University. His research interests include data visualization and geographical visualization. Contact him at rwsimmo1@asu.edu

**Ross Maciejewski** is an assistant professor in the School of Computing, Informatics, and Decision Systems Engineering at Arizona State University. His research interests include geographical visualization and visual analytics focusing on public health, dietary analysis, social media, and criminal incident reports. Maciejewski has a PhD in computer engineering from Purdue University. He is a senior member of IEEE. Contact him at rmacieje@asu.edu.