Visualizing Dynamic Data with Maps

Daisuke Mashima*, Stephen Kobourov†, Yifan Hu‡

*College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
†Department of Computer Science, University of Arizona, Tucson, AZ, USA
‡AT&T Labs – Research, Florham Park, NJ, USA

Abstract—Maps offer a familiar way to present geographic data (continents, countries), and additional information (topography, geology), can be displayed with the help of contours and heat-map overlays. In this paper we consider visualizing large-scale dynamic relational data by taking advantage of the geographic map metaphor. We describe a map-based visualization system which uses animation to convey dynamics in large datasets, and which aims to preserve the viewer's mental map while also offering readable views at all times. Our system is fully functional and has been used to visualize user traffic on the Internet radio station last.fm, as well as TV-viewing patterns from an IPTV service. All map images in this paper are available in high-resolution at [1] as are several movies illustrating the dynamic visualization.

Index Terms—Information Interface and Presentation, Multimedia Information Systems, Dynamic Visualization, Graph Drawing, Spatialization, Map-based Visualization

1 INTRODUCTION

Map representations provide a way to visualize relational data with the help of the geographic map metaphor. Contact graphs, where regions represent nodes and edges are represented by the corresponding regions sharing borders, are an example of such map representations. By definition, contact graphs are limited to planar graphs, but the notion of a map representation can be generalized to non-planar graphs as follows: clusters of well-connected nodes form countries, and countries share borders when neighboring clusters are interconnected. In the context of information visualization, such maps allow us to show not only connectivity information in the underlying data (via nodes and edges between them), but also clustering (via countries). Specifically, by grouping nodes into different colored regions, we can easily see individual clusters and relations between the clusters. Such explicit grouping makes it possible to identify central and peripheral clusters, as well as central and peripheral nodes within clusters. Finally, cut-vertices and edges, often on the border between two clusters, make it clear which nodes and edges connect disparate parts of the data.

While many general users may be put off by overwhelming-looking infographics, maps offer a familiarity which allows us to present complex information. Due to abundance of maps – subway maps, train maps, political maps – most people find them intuitive and non-intimidating. Moreover, familiarity with the traditional way of interacting with digital maps via zooming and panning makes them easy to use. This is one of the reasons why maps offer a promising way for visualizing data. Another reason is that maps tend to encourage viewers to spend time examining them. Preliminary informal experiments indicate that people spend twice as long looking at a map, than at a graph of the same data. While in most computer science applications we like to get things done faster, in this case it is an advantage to have a visualization that is aesthetically appealing and unobtrusively encourages the viewer to spend more time.

There is more than just anecdotal evidence to support the suggestion that maps can be a powerful and effective way to visualize relational data. A recent experiment with user-generated graph layout, collected data from over 70 users of IBM's ManyEyes online data visualization tool, to explore the types of graph layouts that people prefer [30]. The results show that users invariably construct layouts that distinctively group clusters in a spatial region that does not overlap with the spatial region occupied by another cluster. This is accomplished at the expense of stretching some edges, and shrinking others. Moreover, 80% of the users used the edges in a cluster to visually delineate the cluster itself, creating a convex hull around the nodes in the cluster. As the authors of the study point out, although this might seem obvious in retrospect, no automated layout algorithms explicitly attempt to do that. Our map layouts, with clusters represented by regions with clear cluster-hugging borders, achieve both of these user-desired features: explicit clustering and explicit cluster boundaries.

The geographic map metaphor is used for visualizing TV shows and the similarity between them, based on common viewing patterns, in the context of recommendation systems [13]. A companion paper describes the algorithmic details of this map generation approach, and how it can be generalized to arbitrary relational data sets [14]. But in both cases the data under consideration is static.

The problem becomes harder when we would also like to visualize some underlying process. For example, instead of showing a static map of popular TV shows, we would like to see the evolution of this data over the course of one year, and discover which shows become more (or less) popular over that time period. Dynamic map visualization deals with the problem of effectively presenting relationships as they change

over time. Traditionally, dynamic relational data is visualized by animations of node-and-link graphs, in which nodes and edges fade in and out as needed. One of the main problems in dynamic visualization is that of obtaining individually readable layouts for each moment in time, while at the same time preserving the viewer's mental map.

In this paper we explore a new way to visualize dynamic relational data with the help of the geographic map metaphor. We present a functional system that was used to visualize music trends collected from the Internet radio station `last.fm` and TV viewing trends from an IPTV service. Some of the challenges encountered along the way include those related to map layout, as well as those related to animated maps. For map layout, preservation of the viewer's mental map under the dynamics in the data, and readability of each individual layout, are both very important. Issues related to animated maps include effective handling of disappearance (blink and you'll miss it), attention (where to look as the animation is playing), complexity (animated maps try to do too much and end up saying very little), and confidence (viewers of animations are less confident of the knowledge they acquire from animated data than from static data) [16]. To deal with most of these challenges we employ simple and practical solutions. For example: we use a canonical map for stable placement of nodes and labels; we use a clustering guided layout scheme to reduce map fragmentation; we break down changes into intermediate steps so that changes are announced before they happen.

The paper is organized as follows. In Section 2, we give an overview of the static GMap algorithm that we later modify for the dynamic setting. In Section 3, we discuss the algorithmic pipeline to address the dynamic visualization challenges associated with readability and mental map preservation, mainly in the context of Internet radio station trend visualization. In Section 4 we present our prototype implementation in that setting. In Section 5 we describe the use of our system for visualizing TV viewing trends in an IPTV service. In Section 6 we survey related work, and Section 7 concludes the paper with directions for future work.

2 Creating Maps from Graph Data

We begin with a summary of the GMap algorithm for generating maps from static graphs originally proposed in [14]. The input to the algorithm is a relational data set, from which a graph $G = (V,E)$ is extracted. The set of vertices V corresponds to the objects in the data (e.g., artists), and the set of edges E corresponds to the relationship between pairs of objects (e.g., the similarity between a pair of artists). In its full generality, the graph is vertex-weighted and edge-weighted, with vertex weights corresponding to some notion of the importance of a vertex, and edge weights corresponding to some notion of the closeness between a pair of vertices. In the case of music, the importance of a vertex can be determined by the popularity of an artist, derived from the total number of listeners, or by the total number of songs played in a given time period. The weight of an edge can be defined by the strength of the similarity between a pair of artists.

In the first step, a cluster analysis is performed in order to group vertices into clusters, using a modularity-based clustering algorithm [21]. In the second step of GMap, the graph is embedded in the plane using a scalable force-directed algorithm [11] or multidimensional scaling (MDS) [20]. In our approach we use information from the clustering to guide the MDS-based layout. In the third step of GMap, the geographic map corresponding to the data set is created, based on a modified Voronoi diagram of the vertices, which in turn is determined by the clustering and embedding. Here "countries" are created from clusters, and "continents" and "islands" are created from groups of neighboring countries. Borders between countries and at the periphery of continents and islands are created in fractal-like fashion. Finally, colors are assigned to countries, with the goal that no two adjacent countries have colors that are too similar. In the context of visualizing dynamic data, where the relative change of popularity is important, we also use a heat-map overlay to highlight the "hot" regions. Further geographic components can be added to strengthen the map metaphor; for instance, edges can be made semi-transparent or even modified to resemble road networks. In places where there are large empty spaces between vertices in neighboring clusters, lakes, rivers, or mountains can be added, in order to emphasize the separation.

3 Maps of Dynamic Data

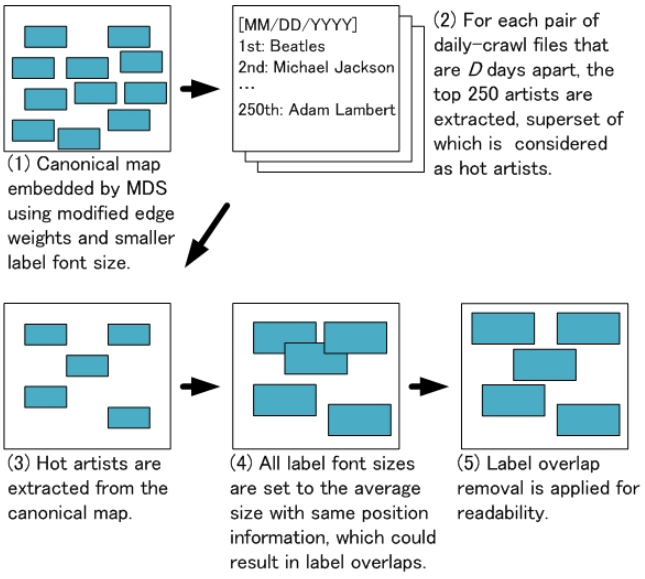


Fig. 1. Algorithmic pipeline to create a base map from a canonical map. (1) Canonical map is created with all crawled artists, by embedding them using MDS and small-sized label fonts. (2) From each pair of daily-crawled files that are D days apart, 250 artists with the highest increase in playcount are extracted. (3) Position data of the hot artists are extracted from the canonical map. (4) All label font sizes are set to the average size. (5) Overlap removal is applied, and the resulting layout is used as a base map.

Static maps of relational data lead to visually appealing representations, which show more than just the underlying vertices and edges. Specifically, by explicitly grouping vertices into different colored regions, viewers of the data can quickly identify clusters and relations between clusters. Moreover, this explicit grouping leads to easy identification of central and peripheral vertices within each cluster.

Extending traditional graph drawing algorithms from static to dynamic graphs is a difficult problem. In most proposed solutions, the typical challenges are those of preserving the mental map of a viewer and ensuring readability of each drawing. Changes are visualized by animation, which can be generated by concatenating static maps, thus providing continuity from one layout to the next. Whereas in dynamic graph drawing it is perfectly reasonable to have vertices move from one moment in time to the next, moving "countries" and "cities" within the countries on a map can be confusing and counter-intuitive. Also, if the layout from one time to the next is significantly different, it is likely that viewers will quickly get lost. A common way to deal with this problem is anchoring some vertices that appear in two or more subsequent drawings. Additionally, the way to encode metrics and changes into the map metaphor needs to be considered. Next we describe how we address some of these challenges. The algorithmic pipeline discussed in Section 3.2 to 3.4 is summarized in Figure 1.

3.1 Last.fm Data

As an Internet radio and music community website, last.fm has over 30 million users. Using a music recommender system, last.fm recommends music based on user profiles. Over several years the recommender system has collected information about how one musician is related to another in terms of how many listeners of one also enjoy the other. For each musician, the last.fm website lists related (similar) musicians. For example, Beethoven is considered to have "super similarity" to Mozart, Bach, Brahms, "very high similarity" to Mendelssohn, Schumann, Vivaldi, and so on. The website also provides the number of listeners of each musician and other metrics related to its popularity. Using this data and daily crawls using the provided API, we create the underlying graph with artists as vertices and with edges determined by the strength of the similarity between the artists at the two endpoints.

3.2 Mental Map Preservation

Mental map preservation is important when visualizing dynamic data. In general, vertices and edges may appear and disappear over time. If a vertex appears, then disappears, and appears again, it would be desirable to use the same location in the layout. Specifically, in the last.fm data artists that were not in the previous map may suddenly become popular while others may drop off from the top.

To address this problem, we create a "canonical map" that stores the position information of a much larger graph than the subgraph that is actually shown. Then, when displaying a specific subgraph consisting of top artists at a given time, we use the pre-computed position information from the canonical map; see Figure 1(1). In this way, as long as the same canonical map is used, the same artists appear in the same position, thereby helping preserve a viewer's mental map. Updating a canonical map is necessary to keep up with trend changes, but it can be done less frequently, for instance once a month, provided that the number of artists included is large enough to ensure that it contains all artists that could appear in the visualization before the next update. Note that such an approach would fail for cases like the stunning debut of Susan Boyle, who catapulted from not-known to top 100 in the span of a few weeks.

3.3 Map Readability

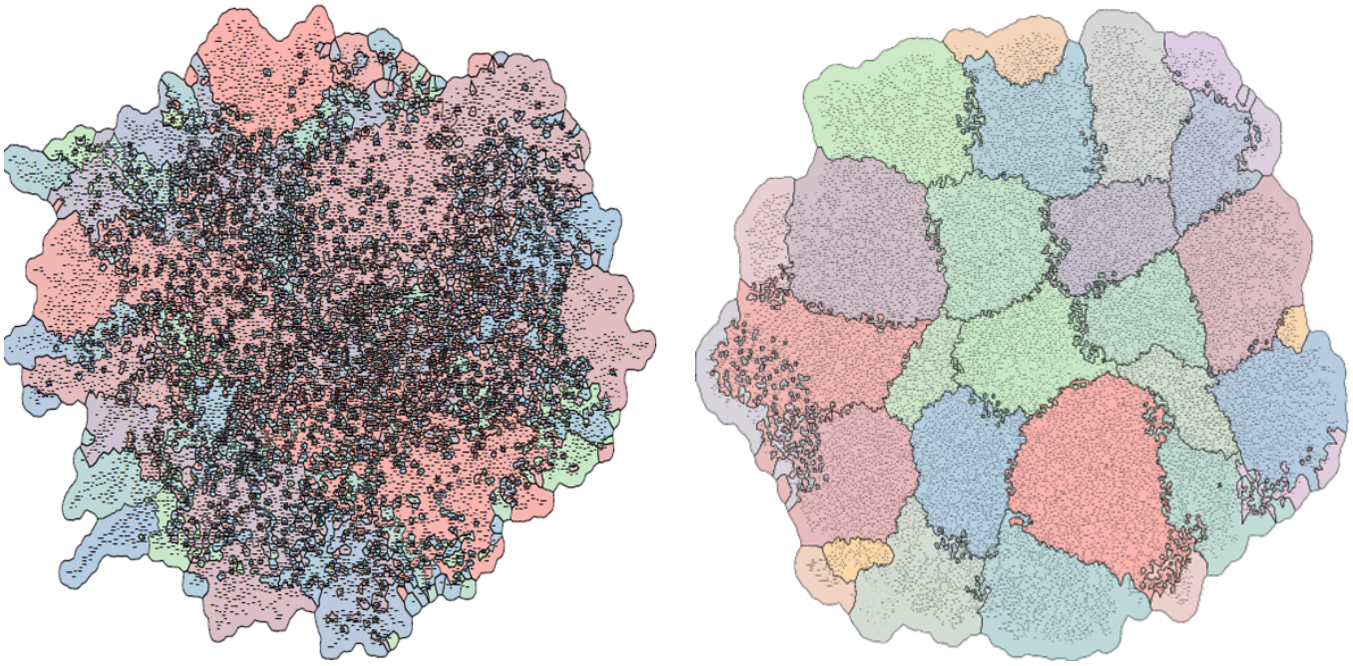
Our initial attempt at obtaining a canonical map with GMap of the 18,000 artists crawled from the top artists in last.fm immediately exposed a problem with this approach. We used modularity and MDS for clustering and embedding, respectively. The (clustering, embedding) pairing seemed applicable, given that in the underlying graph the strength of an edge corresponds to the measure of similarity between the two artists it connects. Since the inverse of similarity can be naturally interpreted as a distance, MDS can determine a layout that matches the underlying clustering.

However, the resulting map was far from ideal; see Figure 2(a). The most conspicuous problem is the fragmentation of countries into disjoint regions. We found that, on average, one cluster (country) is divided into over 100 regions. Even though this canonical map is never intended to be seen by viewers, such fragmentation will negatively affect the readability of resulting visualization. In fact, the placement of the vertices determined by this canonical map led to significant fragmentation even in a map created for the top 500 artists (in terms of the number of listeners). Using a force-directed layout [11] or a LinLog layout [22] in place of MDS resulted in even more fragmentation.

One possibility is that the fragmentation problem is, to some extent, caused by the independent nature of the clustering and the embedding steps. Therefore, we combined the two steps by using the clustering results as additional input parameters of the embedding process. In other words, based on the clustering results, we increase the edge lengths between artists that belong to different clusters, leading to a much better canonical map; see Figure 2(b). In this map, fragmentation is significantly reduced though there are irregularities near some country boundaries. When the canonical map is generated in this way, there is no fragmentation in a map of top 500 artists.

It is worth mentioning that GMap uses a label overlap-removal routine [12] to ensure that vertex labels are readable. This is accomplished by moving apart vertices with overlapping labels, but can potentially lead to a vertex near a border between two countries "jumping" into the wrong country. By strengthening the edges between vertices in the same cluster, we help such vertices stay in their own countries. Even though such edge length modification distorts the underlying raw similarity information, most of the resulting layout changes are local.

Since a smaller number of hot artists will be extracted out of a canonical map in a later step, we need to determine node



(a) (b)

Fig. 2. (a) Map of 18,000 artists (b) Map of 18,000 artists using adjusted edge lengths. Note the significant reduction in cluster fragmentation in the second map.

positions in such a way as to prevent the resulting map from looking too sparse. To accomplish this we simply use smaller label font sizes that are proportional to the popularity of artists; see Figure 1(1).

3.4 Mental Map Preserving Node and Label Placement

A general issue regarding the visualization is the number of artists displayed in a single map. As we target a regular computer screen, the number of artists that can be shown depends on how many non-overlapping and readable labels can fit on it. Not surprisingly maps with 1,000 labeled artists turned out to be unreadable. Cutting down the number of artists to the top 250 leads to better results.

Even within the top 250 artists, some are much more popular than others. One straightforward way is to represent the popularity of an artist by varying the font size of the labels, as in geographic maps where the names of major cities are drawn with larger fonts than those of smaller towns. To modify the font sizes, we use the following conversion for each artist a displayed.

$$ModFontSize_a = BaseSize + Variation * f(pop_a) \quad (1)$$

where

$$f(pop_a) = \frac{pop_a - AVERAGE_{i \in A}(pop_i)}{MAX_{i \in A}(pop_i) - AVERAGE_{i \in A}(pop_i)} \quad (2)$$

Here, the set A denotes all artists to be shown on a map, and pop_a indicates a popularity metric (e.g., the number of listeners) of an artist $a \in A$. Note that $f(pop_a)$ in (2) is scaled to be within $[-1,1]$. Therefore the resulting font size $ModFontSize_a$ is in $[BaseSize - Variation, BaseSize + Variation]$, with a mean font size of $BaseSize$. A sample map created under this configuration is shown in Figure 3. Related to font size modification is the timing of the label overlap-removal step [12]. Because we modify the font sizes after the layout of the nodes in the canonical map has been determined, the resized labels could lead to new overlaps in crowded areas, once again making the maps difficult to read. Applying another overlap-removal step, once the labels have been resized, makes the maps readable but at the expense of modifications in the positions of labels from one time frame to the next. Although this process could be effective for the sake of better presentation, the negative side effects (inconsistent label positions between consecutive frames) seem to outweigh the advantages. While such movements of labels over time would draw viewers' attention to an area where there are changes, movements like these do not fit our general approach for maintaining a viewer's mental map by having a fixed geographic map as a reference.

In order to benefit from overlap removal without moving labels from frame to frame, we adopted the following approach, summarized in Figure 1. First, we create the canonical map. Second, we form the superset of artists that appear on any of the map frames to be included in the animation. Third, we extract the position information for these artists from the canonical map. Fourth, we set the font sizes of all labels on a map to the average size, i.e. $BaseSize$ in formula (1). Fifth, we perform an overlap-removal step and call the final result the "base map" because it is used to create each map frame in the animation. As a result of the pre-processing, the positions of artists and shapes of countries remain unchanged within one animation.

Note, however, that since this base map is generated every time we create an animation, the node positions are not exactly

Fig. 3. The top 250 artists from `last.fm` in July 2009: showing artist popularity through font sizes, while also displaying similarity via proximity in the map.

consistent among animations created for different time periods. For example, today's animation and an animation created one week later could have slightly different node positions and country boundaries owing to both the overlap removal and difference in artists to be displayed. But, when those base maps are created based on the same canonical map, such differences are minimal.

When we evaluated the animations created by the above procedure, we found that the lack of easily recognizable differences among maps can be a problem. Too much of a good thing (mental map preservation) can be bad. Specifically, it is difficult to spot the differences, as only the font sizes of some artists are changing between map frames, while other components (e.g., size and shape of countries) remain exactly the same; a sample animation can be found at [1]. As we would like to keep the mental map of viewers unchanged from one frame to the next, and just changing the font sizes does not convey the changes in the data, we employ another visual cue that is well suited to maps, namely heat-map overlays. We discuss the metric used to create such heat-maps next, and give the detailed procedure for creating them in Section 4.

3.5 Metric for Visualization

A typical challenge in the visualization of dynamic data is defining a suitable metric which allows us to extract "hot" objects (e.g., artists) out of the canonical map and visualizing them meaningfully. Since the suitable metric is highly context-specific, our discussion here focuses on last.fm data and their API. For example, the number of listeners for each artist and the number of times each artist's songs are played (also called playcounts in last.fm) both seem to be useful. However, these numbers are all cumulative. In other words, artists that have been around for a long time tend to have higher values than newer artists who only recently attracted attention. While such numbers are useful to see long-term popularity, it implies that these values largely depend on the past data, and are not significantly affected by recent and short-term dynamics, which are often of interest to the viewers.

Ideally, both long-term and short-term metrics should be incorporated in the visualization. Thus, while using the *cumulative number of listeners as a long-term metric*, we also consider the short-term one, which is more sensitive to abrupt changes. To prevent the bias by past data, we focus on the difference in these values over a fixed time interval. The ideal interval varies depending on the settings and nature of the target data set. In the case of last.fm data, these numbers are updated weekly, so 7-day or longer interval is appropriate. Our preliminary analysis indicates that playcounts capture the dynamics of the moment well, so in our implementation we use *differences in playcounts as a short-term popularity metric*; see Figure 4.

4 Implementation

Using the approach presented in Section 3, here we describe our visualization system applied to last.fm data. In addition, we discuss how to address the four additional challenges that arise in animated cartographic maps, namely disappearance (blink and you'll miss it), attention (where to look as the animation is playing), complexity (animated maps try to do too much and end up saying very little), and confidence (viewers of animations are less confident of the knowledge they acquire from animated data than from static data) [16].

A system implementation overview is shown in Figure 5. The system contains both monthly tasks and daily tasks. The crawling is done using a custom-made Java program and the last.fm API. We use modularity-based clustering [22] and `neato` in `Graphviz` [15] for the MDS-based embedding. To generate animations, we use `ImageMagick` (http://www.imagemagick.org/).

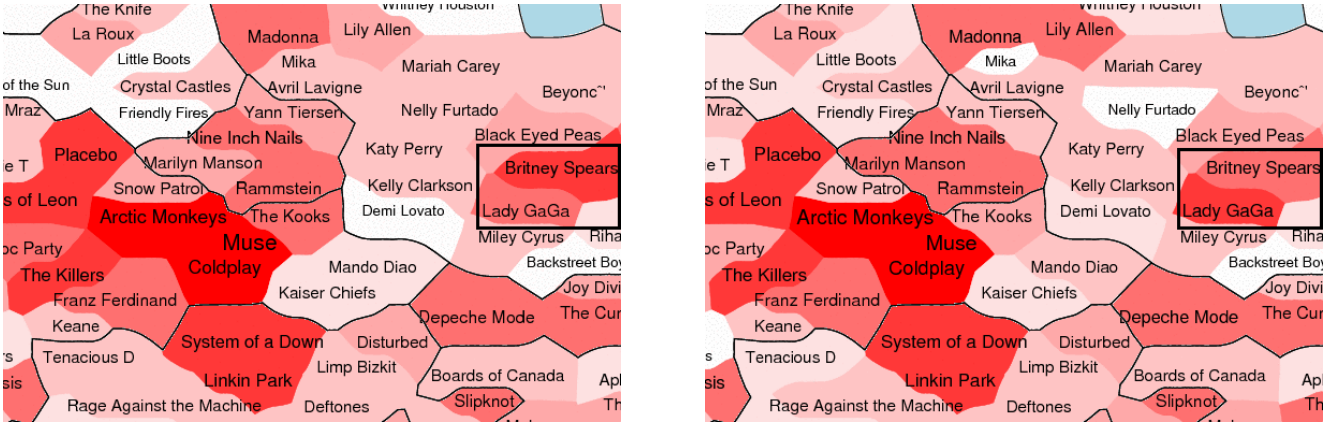


Fig. 4. Two consecutive heat-maps, one month apart in the fall of 2009, showing increased interest in Lady GaGa at the expense of nearby Britney Spears (see areas highlighted by rectangles). The data is from the Internet radio station `last.fm`, which tracks millions of listeners.

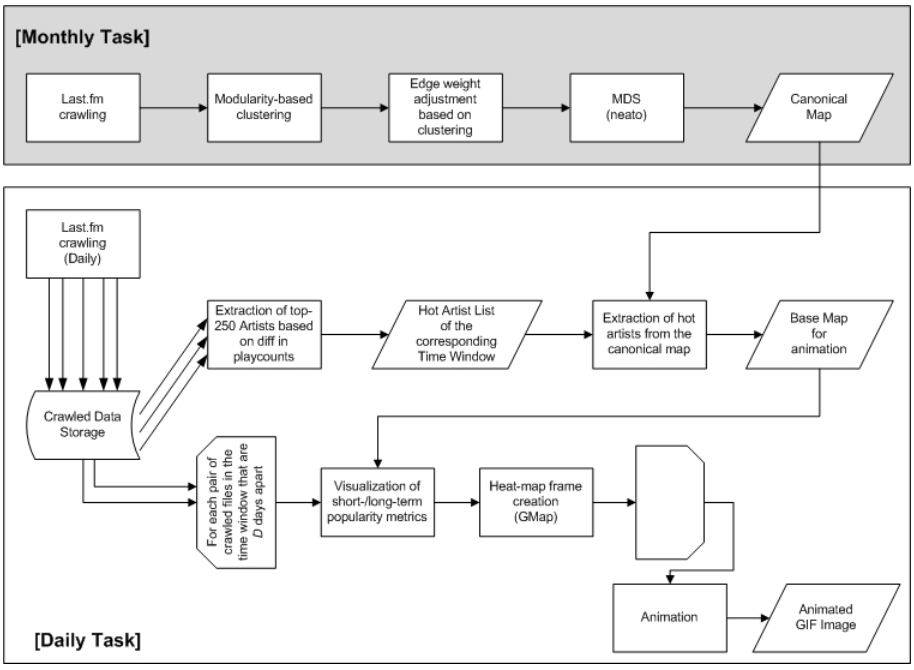


Fig. 5. Overview of Implementation. The tasks are divided into two: monthly tasks to update a canonical map and daily tasks to create heat-map animation.

Canonical Map: creating and updating a canonical map that contains position information of 18,000 artists. Currently, crawling starts with the top-10 artists from the top-50 popular tags on last.fm, and recursively collects information about artists that are similar to them, in a breadth-first fashion. The result is stored in the DOT format used by `Graphviz`, with edge weights defined by the "similarity" values provided by last.fm. This relational data set is fed into the clustering module. Based on this clustering result, we adjust edge lengths as follows in order to reinforce the edges connecting nodes in the same cluster.

1) The length of intra-cluster edges is set to 1.
2) The length of inter-cluster edges is set to a constant $L > 1$ (currently $L = 75$ is used based on the results of our preliminary trials).

Following this step, the graph with adjusted edge lengths is passed on to `neato`, which then computes the node positions for the 18,000 artists. The output from `neato` is used for vertex placement in the canonical map.

Daily Crawling: Daily crawling is done independently and in much the same way as in the monthly task. The results of daily crawls are kept as separate DOT files. It should be possible to make the size of the daily crawl much smaller than the monthly one without missing new and important artists, but we have not explored this as the crawling of 18,000 artists usually completes in less than 12 hours using one PC.

Base Map: selecting daily-crawl results that are in the given time window (one of the configurable parameters in the current implementation) and, for each pair of daily-crawl results with

(a) (b) (c)

Fig. 6. (a) Blue is used to highlight areas where artists are about to disappear: Bon Jovi, Deep Purple, Elvis, Simon & Garfunkel, CCR, and Eric Clapton. (b) Yellow is used to highlight areas where new artists are about to appear. (c) The image after the update, showing newcomers: Bruce Springsteen, Neil Young, The Kinks, and The Beach Boys.

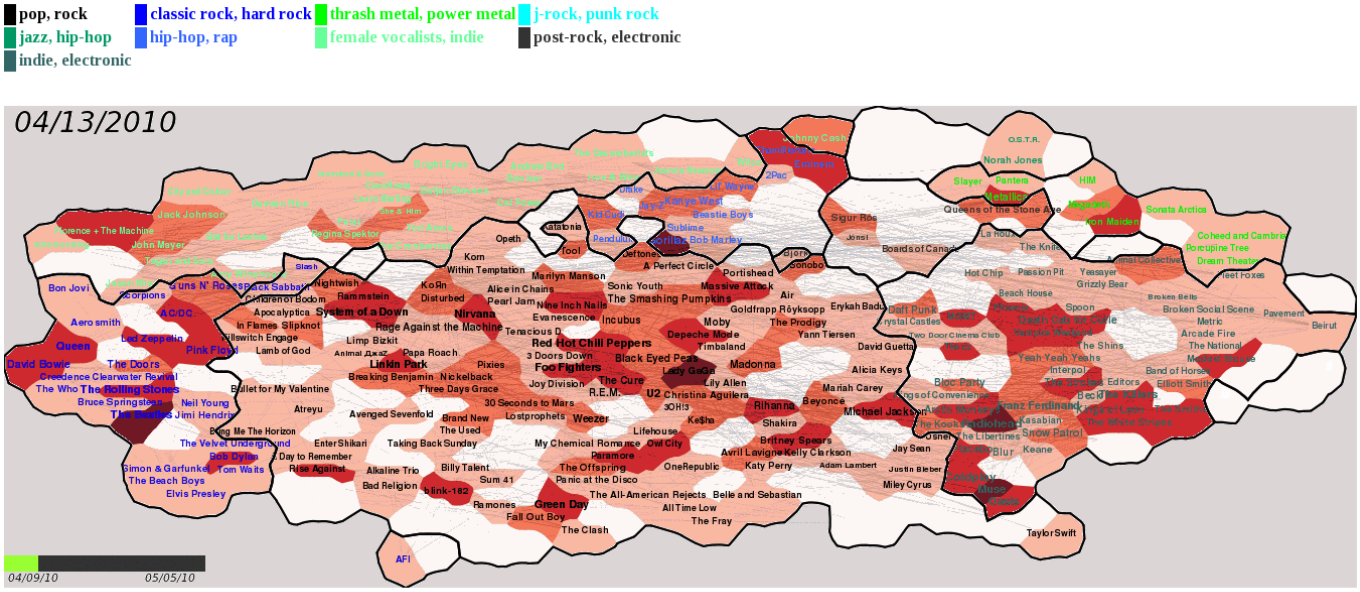


Fig. 7. A snapshot of our trend visualization of last.fm. On the heat-map, darkness of the color indicates the degree of increase in short-term popularity, while font sizes of artists correspond to long-term popularity. Labels of artists who are out of the top 250 are hidden and colored white. Clusters based on similarity are represented with black boundary lines as well as label colors, which are mapped with country names shown on the top of the map. The date label located on the top-left corner indicates the timestamp of the heat-map displayed, and the progress bar on the bottom-left shows the position of the map in the entire animation.

timestamps that are D days apart, computing the differences in playcounts of all the artists. D can also be arbitrarily adjusted based on a characteristic of the target data set or a preferred degree of "sensitivity to changes". Based on these differences in playcounts, the top-250 artists are extracted for each pair of crawled data files, which are D days apart, in the time window. The superset of these artists is saved as a list of "hot" artists. Note that this list could contain more than 250 artists. Since the total number was usually less than 300, the base map remains readable and we include all of them. (As shown later in an example, artists that are not in the top 250 in each frame are hidden.) The position information for nodes in the base map is extracted from the canonical map. As discussed in Section 3.4, we also apply a label overlap-removal step here.

Metrics Visualization: creating an animation that visualizes the changes. Each heat-map frame to be included in an animation is created by modifying font sizes of the base map as well as categorizing artists in the base map into heat-map clusters. While the latter is done based on the magnitude of difference in playcounts, we use the cumulative number of listeners for each artist to determine the font sizes with formula (1). Our implementation uses $0.5 \times BaseSize$ as $Variation$. We also need to establish groups of artists that have similar degree of change in order to draw a heat-map. There are a couple of ways to do this. For example, we can use playcount differences to classify artists, perhaps with suitable log scaling, and map

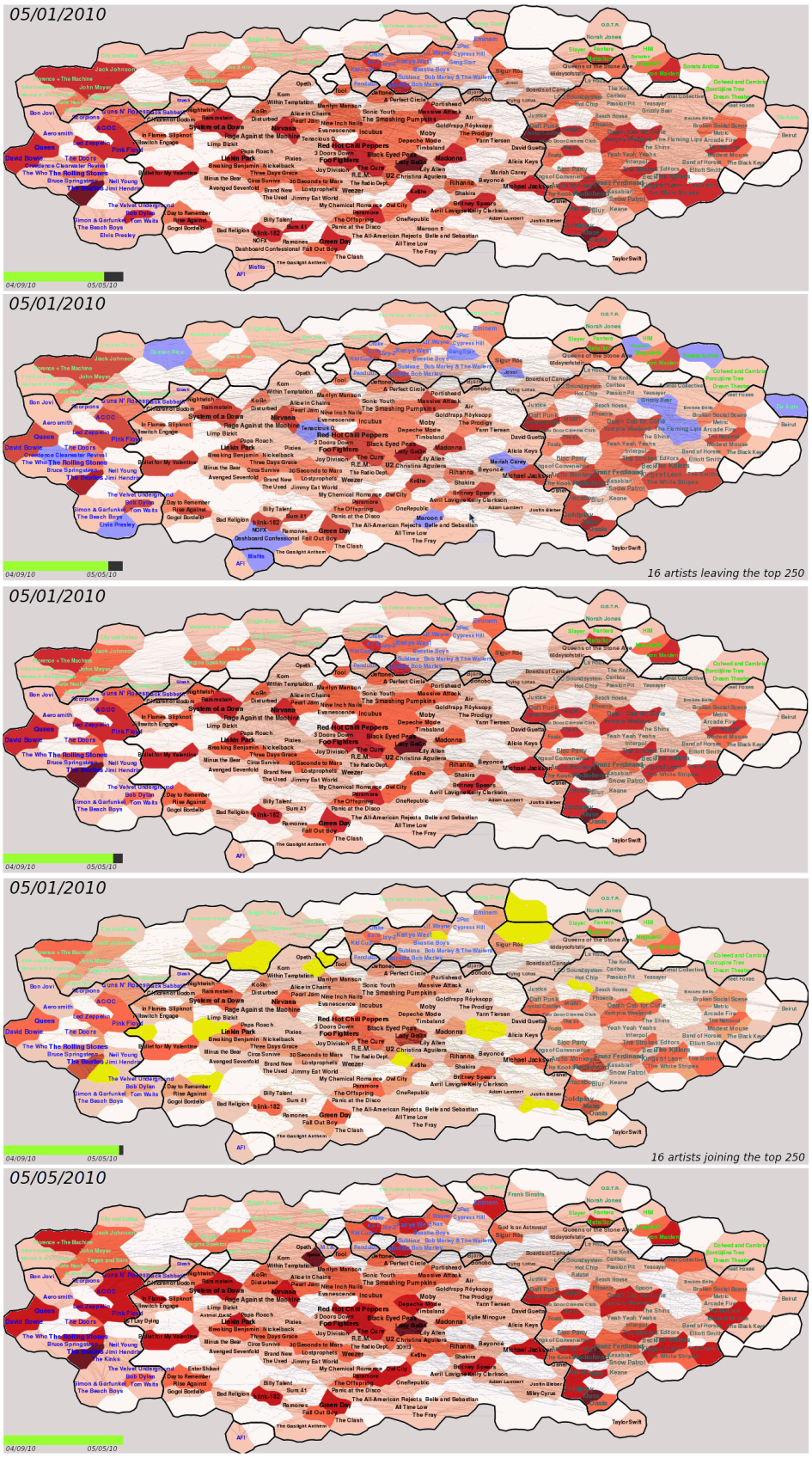


Fig. 8. A sequence of animation frames between two consecutive base heat-maps including blue and yellow highlights.

the scaled differences to a color palette. Alternatively, we can utilize the ranking of artists in terms of the degree of change in playcounts to bin artists, and map bin indices to a color palette. We choose the latter option and use a single-hue color scheme so that artists with larger increase in popularity are assigned darker red colors. In this way, both the overall popularity and the temporal ups and downs of each artist can be visualized in the map.

Countries: when using heat-map overlay, each country can no longer be colored with a uniform color. Even though country boundaries help define the countries, additional visual cues are needed. Otherwise, viewers could have difficulty in identifying similarity relationships when countries are fragmented. We use the original clustering information based on similarity to define a label color for each artist so that artists in the same country have the same color. Currently we simply select a label color from a static palette, but this can be improved, for example by using a maximal differential color scheme [18], which is part of our future work. Within a country, different shades of the background color indicate variations in popularity. Each country is also automatically "named"; these names are created by taking the top two most frequent tags assigned to artists in the corresponding country. A list of country names written with the color associated with that country is presented above the animation; see Figure 7. In this way we obtain a labeled heat-map image, or the "base heat-map", and this process is repeated for each pair of daily-crawl results in the target time window that are D days apart.

Attention, (Dis)appearance, Complexity, and Confidence: The base heat-map images are concatenated in chronological order to generate a single animated GIF file. Note that we can arbitrarily change the number of frames in an animation by adjusting the size of the time window accordingly. However, naive concatenation of image files would create an animation that is difficult for a viewer to follow because there are too many changes happening at the same time: artists disappearing/appearing in the top 250, in addition to heat-map color changes representing artists remaining in the top 250 but whose popularity has changed. We break down these changes in a few intermediate steps. To emphasize appearance and disappearance of artists from one frame to another, we create intermediate frames for each pair of base heat-maps as follows.

1) A frame that highlights in blue all disappearing artists; see Figure 6(a).
2) A frame that hides all disappearing artists and updates heat-map colors for artists decreasing their popularity.
3) A frame that highlights in yellow all appearing artists; see Figure 6(b).
4) A frame that shows the artists joining the top 250 and updates heat-map colors for artists increasing their popularity, which is also the next base heat-map; see Figure 6(c).

Thus, frames 1 and 2 correspond to a "disappearing/decreasing" phase and frame 3 and 4 establish an "appearing/increasing" phase. To help viewers understand which part of the animation they are watching, we also include a date label and a progress bar.

A snapshot of our last.fm visualization, which incorporates all the components discussed so far, is shown in Figure 7. Figure 8 shows a sequence of animation frames between two consecutive base heat-maps (for May 1, 2010 and May 5, 2010). An animated version is also available online at [1]. The current implementation of our system accepts the entire duration of the animation and the number of heat-map frames as configurable parameters, and the interval between frames (4 days in this example) is determined from these parameters. As can be seen in this example, ups and downs in short-term popularity can be easily recognized by comparing darkness of colors. Artists joining or dropping off from the top 250 are highlighted, and artists out of the top 250 are hidden. In addition, changes between frames are two-phased which helps viewers identify and keep track of differences. Detailed evaluation of the "complexity" of these maps requires a user study. However, we believe that the metrics are reasonably encoded in familiar map components and that the geographic map metaphor helps viewers intuitively understand them. For example, short-term changes in playcounts are encoded with color changes, while long-term popularity changes are encoded with label font sizes. The combination of these long-term and short-term metrics allows the viewer to derive additional information, such as spotting a rising star that suddenly draws public attention. Country boundaries and label colors help the viewer grasp the relationships between similar artists at a glance. More detailed similarity information is also conveyed by semi-transparent edges coming from the underlying graph.

One of the best ways to increase viewer "confidence", when looking at animated maps, is to give the viewer the ability to pause, rewind, and replay the animation. Because this is difficult to achieve with a GIF animation, we also provide movies which offer better control via movie-player interface (also available at [1]).

After a brief review of the animation used to illustrate this paper, it is easy to identify some trends and patterns. Several artists, such as the Beatles, Lady GaGa, and Radiohead, are popular throughout the time period. Others, such as Michael Jackson, fluctuate in popularity but remain in the top 250. Still others, such as Adam Lambert from American Idol, go in and out of the top 250. We can also spot some events in music industry. For example, the release of the new album "Fever" by Bullet For My Valentine on April 27 is accompanied by a characteristic color change: the band was pink through much of April, but suddenly surged to dark red on April 27 and remained so thereafter.

Our system takes 3-4 hours to generate a canonical map from a crawled data of 18,000 artists when run on a single machine. This is mainly due to the embedding step using MDS. However, this process is done infrequently (monthly in our implementation) and is still quick enough to allow for daily updates. The daily tasks, including creation of a base map and a heat-map animation, require less than 15 minutes for the creation of an animation of 1-month duration. Therefore, our visualization scheme can support daily update.

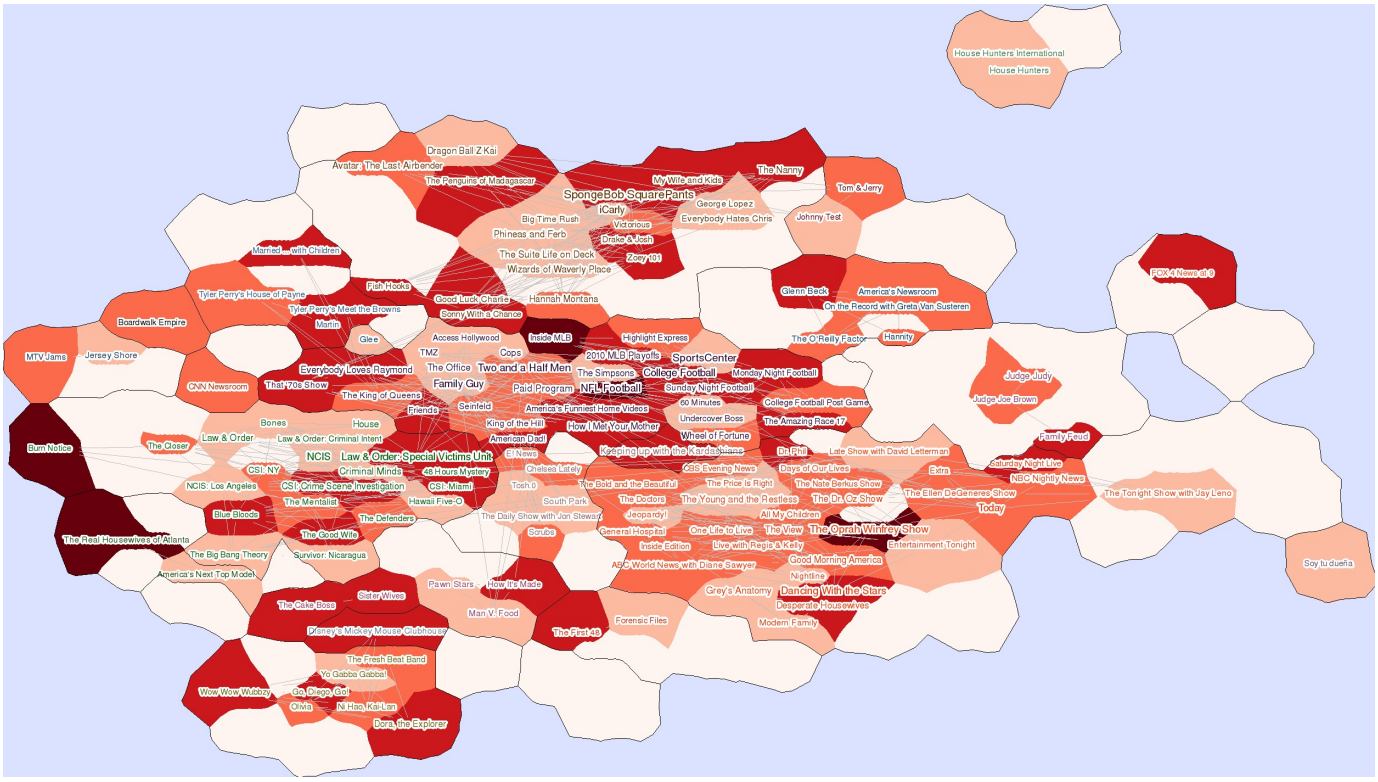


Fig. 9. A heat-map visualization of the IPTV dataset. The label font sizes represent popularity for each 1-month time period, whereas darkness of heat-map colors corresponds to the magnitude of difference in popularity from the previous snapshot. For instance, "Inside MLB," shown with relatively small label but with the darkest heat-map color, is a program that gained popularity in a short time period, while "NFL Football" is a program with constantly high popularity.

5 Application to IPTV Data

In order to demonstrate the wide applicability of our scheme, we customized the system described so far to visualize data obtained from U-verse, which is an IPTV service offered by AT&T. The data includes popularity of the top 1,000 popular TV programs for each 1-month period, and, for each program, the top 10 most similar TV programs[1]. Similarity between programs are calculated as in [13].

Our dataset therefore contains names of TV programs and popularity metrics, which we use as nodes with weights, and similarity relations, which we can use as edges. Due to the differences in some parameters of this dataset compared to the last.fm dataset, we tailored the system as follows. First, we use only one of the datasets, made of 1,000 TV programs, to create a canonical map. Even though the number might look smaller than the last.fm visualization, this choice can be justified because the line-up of TV programs does not change significantly during a TV season. Second, for the IPTV data sets, we only have one metric indicating popularity. Thus we directly used the metric to determine the label font sizes of each program while using difference from the previous data set to determine heat-map cluster colors. This way, the resulting visualization can convey overall popularity and momentum at the same time. For example, programs with small label text and dark red color, such as "Inside MLB" in Figure 9, are those that quickly attracted viewers. Additionally, because TV program names tend to be long, we use relatively smaller fonts. To make the labels easier to see, we adopted a technique used in map making, where a white background similar in outline to the label, but slightly larger, is rendered before the label.

We found that the change in ranking based on popularity is more dynamic in our IPTV dataset than in our last.fm dataset. If we use the top 250 programs when generating a base map as discussed in Section 4, the number of labeled nodes in the base map becomes too large to fit a regular computer screen. Thus, we decided to use the top 150 programs of each interval as "hot" TV programs of the corresponding period. Another difference from the last.fm visualization is the lack of label legends because our datasets did not contain parameters that could be used to automatically generate such legends (such as tags on last.fm). The color of the label still indicates cluster membership, and often the nature of nodes in a cluster can be used to determine its type.

Figure 10 shows a sequence of animation frames. An online animated version can be found at [1]. For this IPTV visualization, we used 7 data sets, which were collected for 9/15/2010 - 10/15/2010, 10/16/2010 - 11/15/2010, 11/15/2010 - 12/15/ 2010, 12/16/2010 - 1/15/2011, 1/16/2011 - 2/15/2011, 2/13/2010 - 3/15/2011, and 3/16/2010 - 4/15/2011 respectively. They roughly correspond to the TV season starting on September, 2010.

From Figure 10, we can observe several patterns and trends.

1. All the data was collected in accordance with appropriate end user agreements and privacy policies. The analysis was done with data that was aggregated and fully anonymized. No personally identifiable information was collected in connection with this research.

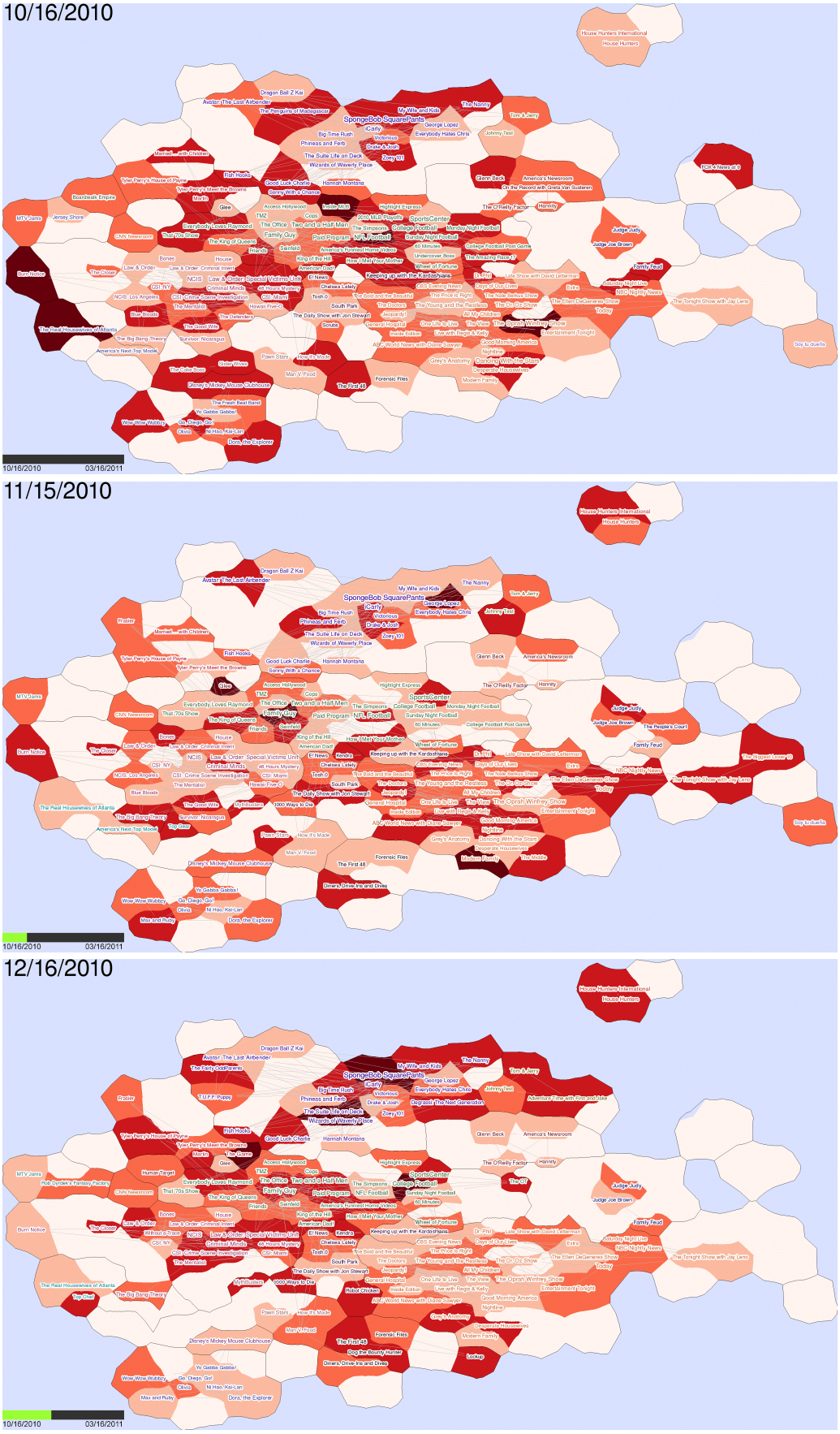


Fig. 10. Frames from the animated heat-map visualization of trends in IPTV. A number of observations can be made. For example, "Burn Notice," located in left of the map, is initially shown with the darkest color but gradually loses popularity and eventually disappears from the map. This corresponds to the fact that the show began in June, 2010 and ended in mid-December that year. (continue on the next page).

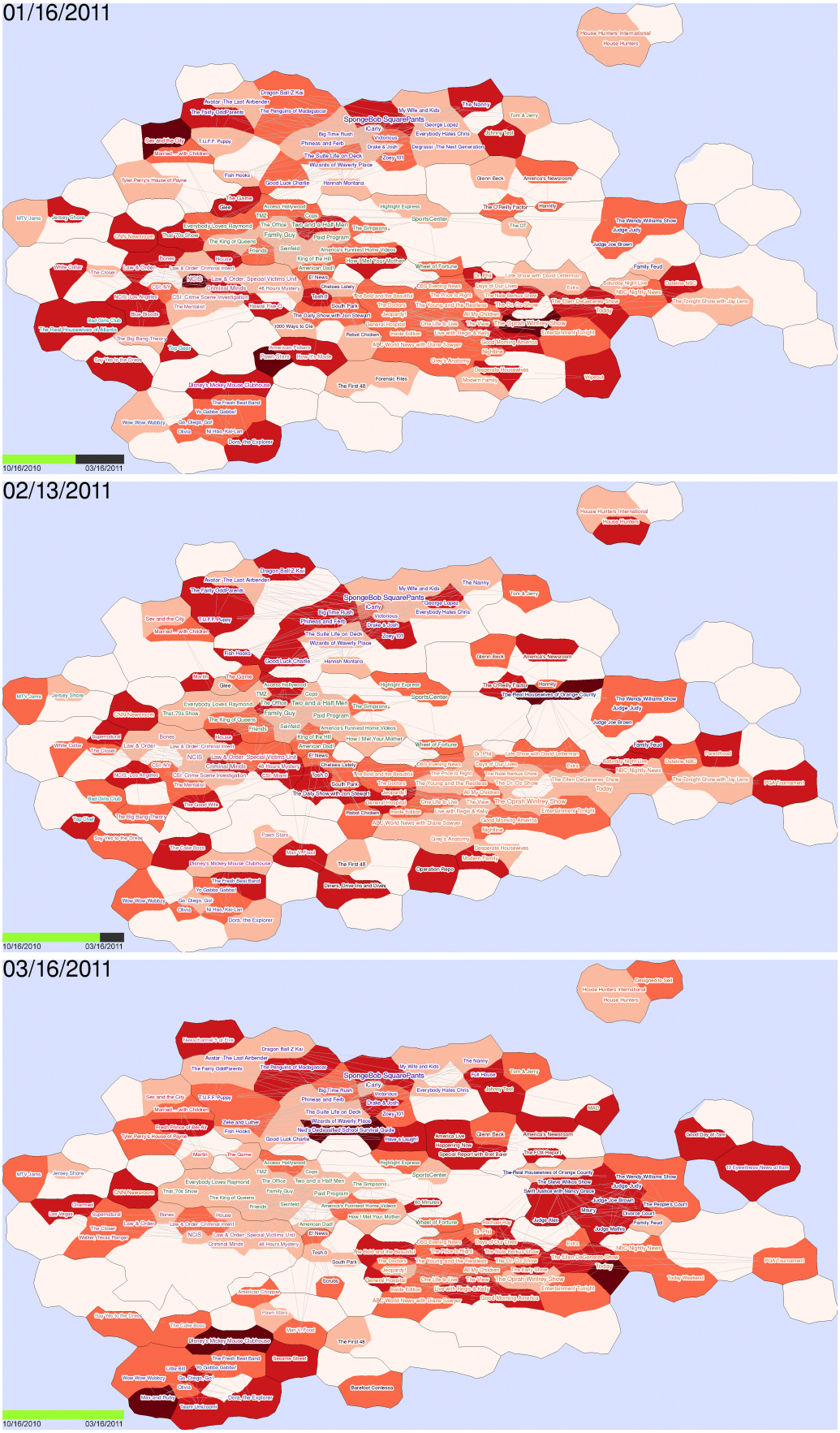


Fig. 10. (continued) Another interesting finding is that “Sex and The City” near the top-left corner suddenly appears in the darkest color in January, 2011 and remains on the map until March, which might have been the result of the Golden Raspberry Award.

For example, "Burn Notice" is initially shown with the darkest color but gradually loses popularity and eventually disappears from the map. This corresponds to the fact that the show began in June, 2010 and ended in mid-December that year. Near the center of the map, we can find that "Inside MLB" and "2010 MLB Playoffs" are popular in October, 2010, but disappears together in November and never appears again, a pattern associated with the length of the MLB season. Another interesting finding is that "Sex and The City" suddenly appears in the darkest color in January, 2011, which might have been the result of the infamous Golden Raspberry Awards.

6 Related Work

The problem of drawing dynamic graphs is well studied; see the survey paper by Branke [5]. In dynamic graph drawing the goal is to maintain a nice layout of a graph that is modified via operations such as inserting/deleting edges and inserting/deleting vertices [5]. Techniques based on static layouts have been used [17]. North [23] studies the incremental graph drawing problem in the DynaDAG system. Brandes and Wagner adapt the force-directed model to dynamic graphs using a Bayesian framework [4]. Diehl and Görg [9] consider graphs in a sequence to create smoother transitions. Brandes and Corman [3] present a system for visualizing network evolution in which each modification is shown in a separate layer of a 3D representation with vertices common to two layers represented as columns connecting the layers. Thus, mental map preservation is achieved by pre-computing good locations for the vertices and fixing the position throughout the layers. Animations as a mean to convey an evolving underlying graph have also been used in the context of software evolution [7] and scientific literature visualization [10].

High-dimensional data is usually visualized as a collection of points in 2-dimensional space using principal component analysis [19], multidimensional scaling [20], force directed algorithms [11], or non-linear dimensionality reduction [26], [29]. These embedding algorithms tend to put similar items next to each other.

There have been several earlier efforts to visualize the Internet radio station last.fm. Graph-based representations have been used [2], with each artist as a node, similarity relationships denoted by edges, and tags used for grouping and coloring. Even though this visualization contains a good amount of information, such as popularity, similarity, tags, and so on, it suffers in readability due to significant node overlapping and fragmentation of groups. Another last.fm visualization [24] uses self-organizing maps that leads to a 2D grid layout in which similar bands are close to each other. but this approach has high computational complexity and does not scale well to large data sets.

Using maps to visualize non-cartographic data has been considered in the context of spatialization [28]. Map-like visualization using layers and terrains to represent text document corpora dates back to 1995 [31]. The problem of effectively conveying change over time using a map-based visualization was studied by Harrower [16].

Also related is work on visualizing subsets of a set of items. Areas of interest in a UML diagram can be highlighted using a deformed convex hull [6]. Isocontours-based bubblesets can be used to depict multiple relations defined on a set of items [8]. Automatic Euler diagrams, which show the grouping of subsets of items by drawing contiguous regions around them have also been considered [27]. Apart from differences in the algorithms used to generate regions, all of these approaches differ from ours in that they create regions that overlap with each other, whereas we take the map metaphor strictly and assume that regions do not overlap.

Robertson *et al.* [25] evaluate the effectiveness of three trend visualization techniques. The results indicate that animation is not well suited to data analysis, but it is often enjoyable and exciting. Since one of the main goals of our work is to create an appealing and informative visualization for the general public, and not a precise data analysis tool, we believe our use of animation for trend visualization is justified. We addressed some of the main shortcomings of animations with the help of strong mental map preservation and a familiar geographic map metaphor.

7 Conclusion and Future Work

In this paper we explored a way to visualize large-scale dynamic relational data with the help of the geographic map metaphor. We addressed some challenges created by the dynamics in the data and presented a system that visualizes the user traffic on the Internet radio station last.fm with a heat-map animation. We believe the applicability of our approach is not limited to the last.fm data. For example, our scheme can be used, with minor modification in the data collection module, to visualize trends in the popularity of web sites, TV shows, etc., where similarity and popularity information are easy to define. The feasibility of such application is demonstrated by our IPTV data visualization.

The major component of our future work is the evaluation of the effectiveness of our visualization through the user study. This would also include the calibration of parameters (duration of animation, interval for difference calculation, etc.). We are also working on implementing a functional interactive interface. As the underlying data is a map, we are exploring pan-and-zoom Google-Maps-like interactions. In this direction, we have developed a searchable interface for static maps which allows users to zoom in and out of our maps and explore by means of intuitive mouse operations; see Figure 11. The integration of our trend visualization into such an interface is part of our future work. The resulting system can be further enhanced by allowing access to external online content (e.g., accessing the last.fm artist web pages, web sites of TV programs, or wikipedia pages) by clicking on node labels. Moreover, although our prototype uses animated GIF, we plan to explore other file-size efficient ways of creating the animation. Finally, offering several metrics for visualization would result in a more powerful system; for instance, our system uses a difference in the number of times each artist's songs are played (playcounts), while the second-order difference in playcounts will allow for more precise view of the momentum of an artist.

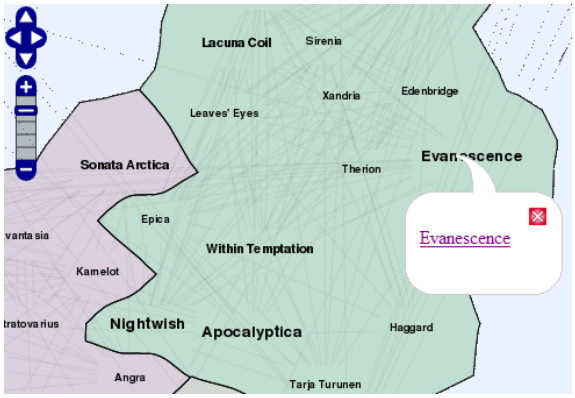


Fig. 11. Map representation with interactive interface

Acknowledgment. This work began in 2009, when all three authors worked at AT&T. We thank anonymous reviewers for the conference version of this paper for very useful comments and suggestions for improvement.

REFERENCES

[1] http://www2.research.att.com/~yifanhu/TrendMap/.
[2] Reconstructing the structure of the world-wide music scene with last.fm. http://sixdegrees.hu/last.fm/index.html.
[3] U. Brandes and S. R. Corman. Visual unrolling of network evolution and the analysis of dynamic discourse. In *IEEE INFOVIS'02*, pages 145–151, 2002.
[4] U. Brandes and D. Wagner. A Bayesian paradigm for dynamic graph layout. In *5th Symp. on Graph Drawing (GD)*, pages 236–247, 1998.
[5] J. Branke. Dynamic graph drawing. *Drawing graphs*, 2025:228–246, 2001.
[6] H. Byelas and A. Telea. Visualization of areas of interest in software architecture diagrams. In *ACM SoftVis'06*, pages 105–114, 2006.
[7] C. Collberg, S. G. Kobourov, J. Nagra, J. Pitts, and K. Wampler. A system for graph-based visualization of the evolution of software. In *ACM SoftVis'03*, pages 77–86, 2003.
[8] C. Collins, G. Penn, and S. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE TVCG*, 15(6):1009–1016, 2009.
[9] S. Diehl and C. Görg. Graphs, they are changing. In *10th Symp. on Graph Drawing (GD)*, pages 23–30, 2002.
[10] C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee. GraphAEL: Graph animations with evolving layouts. In *11th Symp. on Graph Drawing (GD)*, pages 98–110, 2003.
[11] T. Fruchterman and E. Reingold. Graph drawing by force directed placement. *Software-Practice and Experience*, 21:1129–1164, 1991.
[12] E. R. Gansner and Y. F. Hu. Efficient node overlap removal using a proximity stress model. In *16th Symp. on Graph Drawing (GD)*, volume 5417, pages 206–217, 2008.
[13] E. R. Gansner, Y. F. Hu, S. Kobourov, and C. Volinsky. Putting recommendations on the map: visualizing clusters and relations. In *3rd ACM Conf. on Recommender Systems (RecSys)*, pages 345–348, 2009.
[14] E. R. Gansner, Y. F. Hu, and S. G. Kobourov. GMap: Visualizing graphs and clusters as maps. In *IEEE Pacific Visualization Symp. (PacVis)*, pages 201–208, 2010.
[15] E. R. Gansner and S. North. An open graph visualization system and its applications to software engineering. *Software - Practice & Experience*, 30:1203–1233, 2000.
[16] M. Harrower. Tips for designing effective animated maps. *Cartographic Perspectives*, 44:63–65, 2003.
[17] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.
[18] Y. F. Hu, S. Kobourov, and S. Veeramoni. On maximum differential graph coloring. In *18th Symp. on Graph Drawing (GD)*, 2010.
[19] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.
[20] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Press, 1978.
[21] M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582, 2006.
[22] A. Noack. Energy-based clustering of graphs with nonuniform degrees. In *13th Symp. on Graph Drawing (GD)*, pages 309–320, 2005.
[23] S. C. North. Incremental layout in DynaDAG. In *4th Symp. on Graph Drawing (GD)*, volume 1027, pages 409–418, 1996.
[24] E. Pampalk. Islands of music - analysis, organization, and visualization of music archives. *Journal of the Austrian Society for Artificial Intelligence*, 22(4):20–23, 2003.
[25] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14:1325–1332, 2008.
[26] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.
[27] P. Simonetto, D. Auber, and D. Archambault. Fully automatic visualisation of overlapping sets. *Computer Graphics Forum*, 28:967–974, 2009.
[28] A. Skupin and S. I. Fabrikant. Spatialization methods: a cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science*, 30:95–119, 2003.
[29] J. B. Tenenbaum, V. V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.
[30] F. van Ham and B. E. Rogowitz. Perceptual Organization in User-Generated Graph Layouts. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 14(6), NOV 2008.
[31] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *IEEE Symp. on Information Visualization*, pages 51–58, 1995.

Daisuke Mashima is a Ph.D student at Georgia Institute of Technology. He received BE and MS degrees in Engineering from Keio University, Japan. After that, he worked as a software engineer for NTT Advanced Technology Co. His primary research focus is information security, which also covers user-friendly data visualization to help users recognize abnormal situations such as attacks and identity theft cases.

Stephen G. Kobourov received BS degrees in Computer Science and Mathematics from Dartmouth College (1995), an MS degree in Computer Science at Johns Hopkins University (1997) and a PhD degree in Computer Science from Johns Hopkins University (2000). He is an Associate Professor of Computer Science at the University of Arizona. His primary research interests are in geometric algorithms, graph drawing, and information visualization.

Yifan Hu Yifan Hu is a principal member of technical staff in the Information Visualization Department at AT&T Labs - Research. He received a BS and MS in Applied Mathematics from Shanghai Jiao-Tong University (1981, 1985) and Ph.D. in Optimization from Loughborough University, UK (1992). His research interests include numerical and combinatorial algorithms, information visualization and data mining.