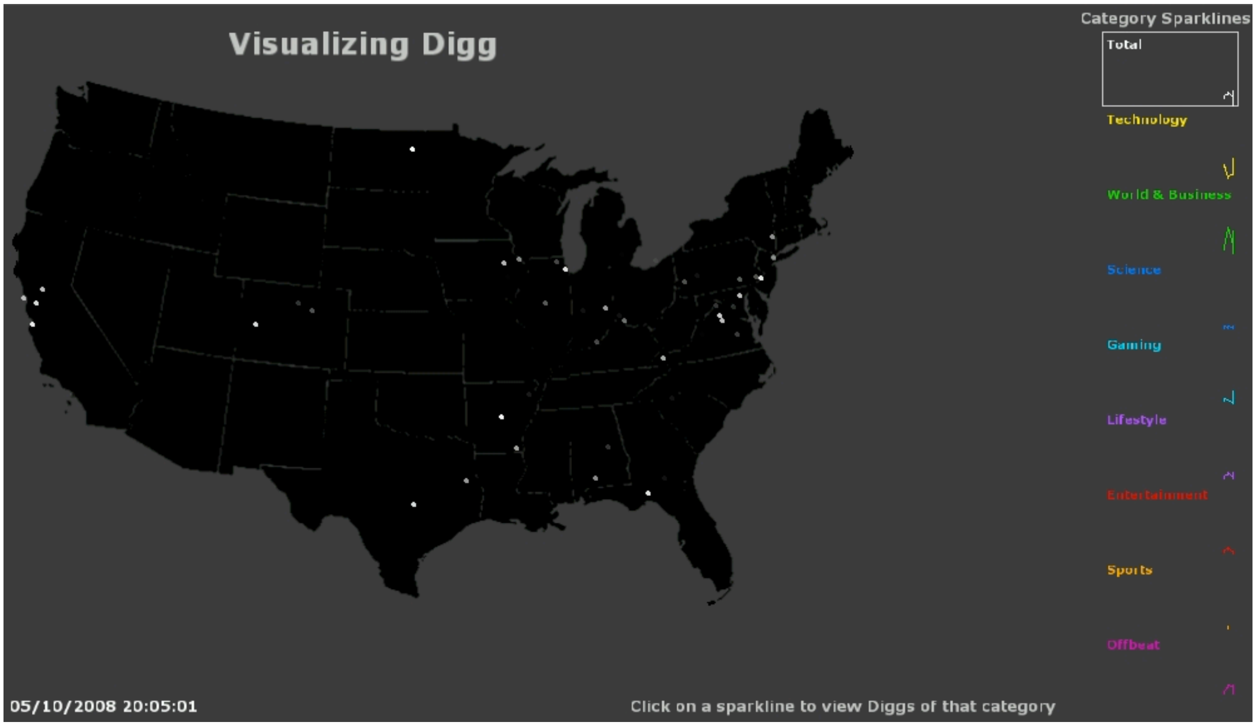
Diane Ko
Ashkan Soltani
i247 – Final Project Synopsis
05/07/2008



Click on a sparkline to view Diggs of that category

http://tekgen.com/dv/applet_v1/

1. PROJECT GOALS

Our project attempts to visualize user activity on a news related website in an attempt to give a sense of human activity on the web. Stories on Digg.com are classified into of seven categories and then promoted by users of the site. We'd like to explore patterns in location, category, and time for this activity.

This visualization is intended to serve as more of an ambient display rather than one specific to data analysis. Dots on a map representing Diggs and sparklines on the side corresponding to the seven categories attempt to convey a sense of how topical activity flows with time and space.

2. RELATED WORK

2.1 Flight Patterns

The main inspiration for our visualization comes from Aaron Kobin's 'Flight Patterns'. This visualization tracks the movement of airline flights across the United States over a pre-set period of time. Our visualization displays Diggs over time in a similar fashion although our use of time is slightly different. We display past Diggs at an accelerated rate then proceed to show live activity in real time once the visualization has 'caught up' to the current time. [1]

FLIGHT PATTERNS

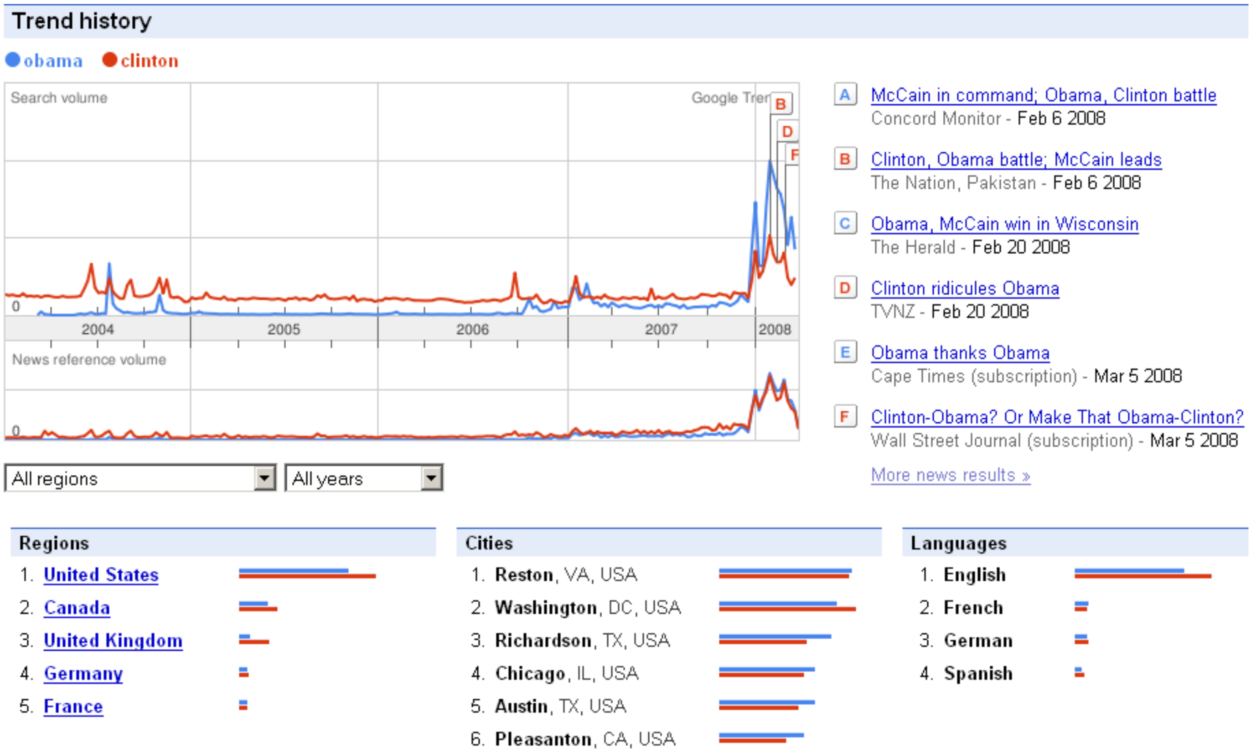
Flight Patterns[2]

[1] To our dismay, we discovered that the 'Flight Patterns' visualization was actually not done in real-time, but was instead generated offline and composited using post-production video rendering tools. This posed as an interesting challenge to replicate in real-time using processing.

[2] http://www.aaronkoblin.com/work/flightpatterns/

2.2 Google Trends

There have been a variety of visualizations attempting to show trends in activity on the web. Most notable is Google Trends, which provides real time statistics on "hot" topics over certain regions. Trend topics are visible in a line graph with news headlines attached to certain regions of the graph. This visualization uses search volume on Google as the basis for the line graphs whereas our sparklines are based on the volume of Diggs per category.



A [McCain in command; Obama, Clinton battle](#)
Concord Monitor - Feb 6 2008

B [Clinton, Obama battle; McCain leads](#)
The Nation, Pakistan - Feb 6 2008

C [Obama, McCain win in Wisconsin](#)
The Herald - Feb 20 2008

D [Clinton ridicules Obama](#)
TVNZ - Feb 20 2008

E [Obama thanks Obama](#)
Cape Times (subscription) - Mar 5 2008

F [Clinton-Obama? Or Make That Obama-Clinton?](#)
Wall Street Journal (subscription) - Mar 5 2008

More news results »

Regions
1. United States
2. Canada
3. United Kingdom
4. Germany
5. France

Cities
1. Reston, VA, USA
2. Washington, DC, USA
3. Richardson, TX, USA
4. Chicago, IL, USA
5. Austin, TX, USA
6. Pleasanton, CA, USA

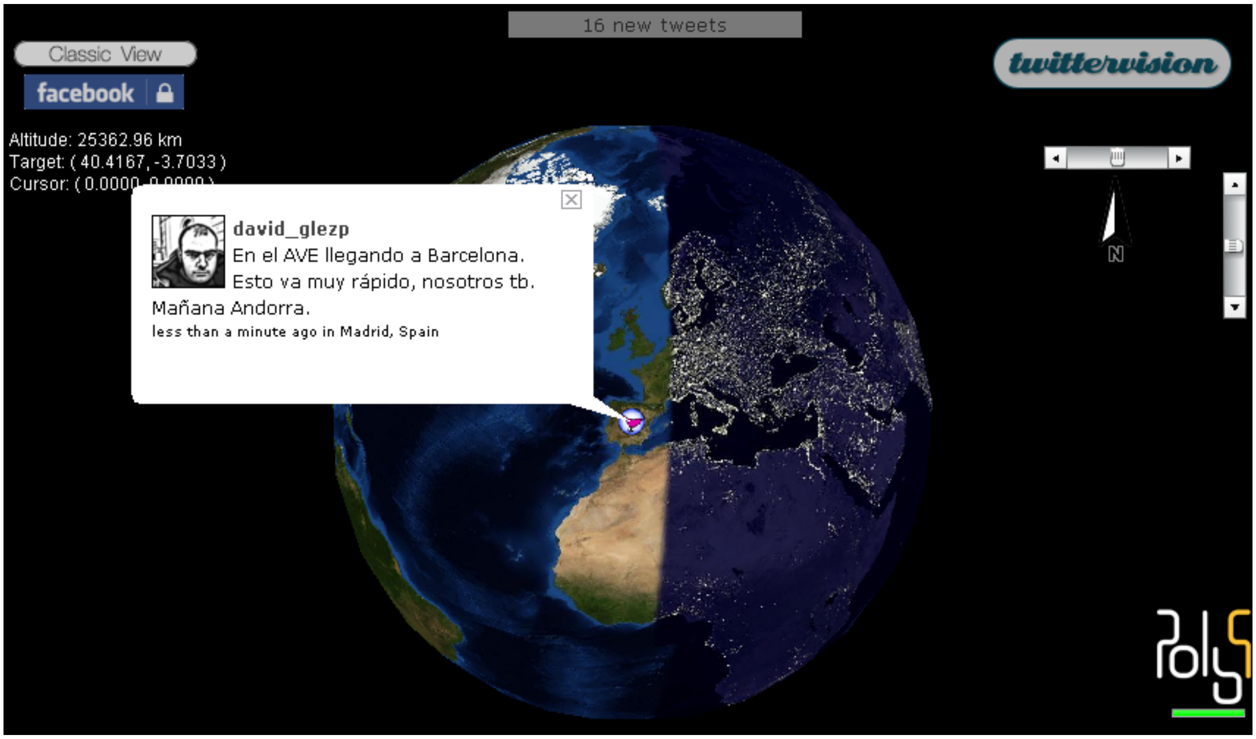
Languages
1. English
2. French
3. German
4. Spanish

Google Trends[3]

[3] http://www.google.com/trends

2.3 Twittervision

Twittervision uses Twitter data to place Twitter posts in the originating location with the creating user's data visible. As each new Twitter post becomes available, the globe rotates to the appropriate global position to show where the post originates from. The globe also shows a real time indication of night and day on the globe. Our visualization focuses less on what is actually posted and more on category of the content and location of the poster. Unlike Twittervision, our visualization focuses only on the United States and the map itself does not move.



david_glezp
En el AVE llegando a Barcelona. Esto va muy rápido, nosotros tb. Mañana Andorra.
less than a minute ago in Madrid, Spain

Twittervision[4]

[4] http://twittervision.com/maps/show_3d

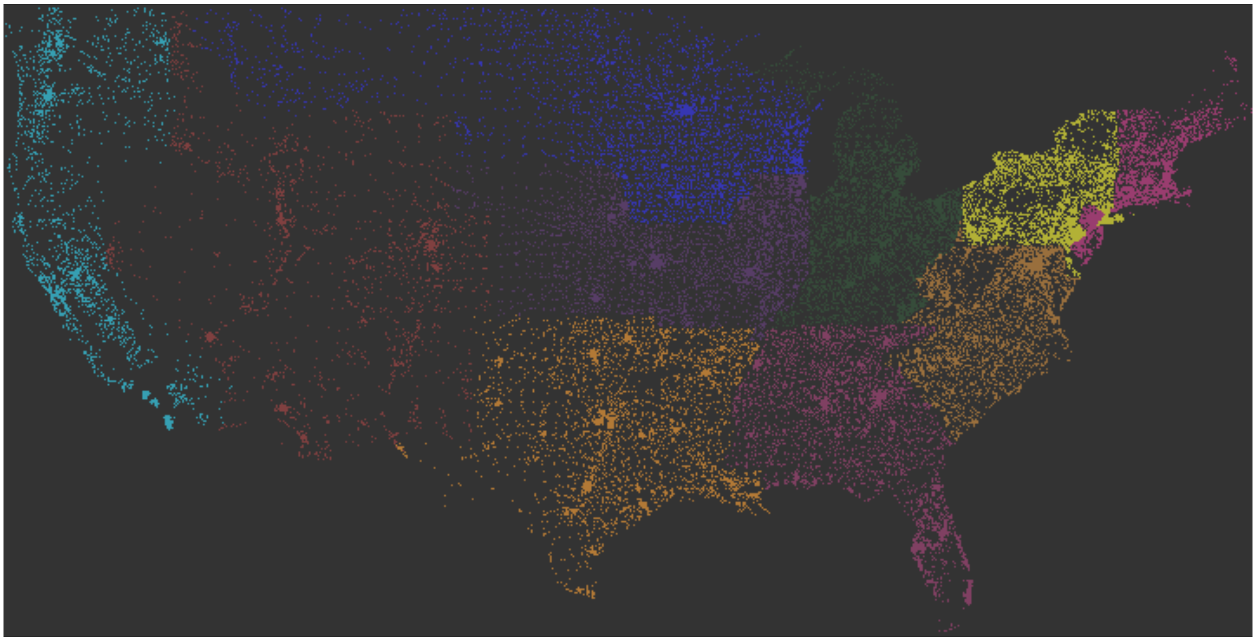
3. DESIGN DECISIONS

Our visualization can be broken down into two main parts: the map and the sparklines. The map is a simple cartogram that shows the actual location of where people Digg from. The category sparklines indicate the volume of Diggs in each category.

3.1 Map

3.1.1 Category Activity by Region

In our initial design, we divided the map into 10 regions in accordance with the US zipcode allocation. We then attempted to use color to indicate the most popular topic or category for that region.



Initial map design: regions based off of first zip code number colored by majority category.

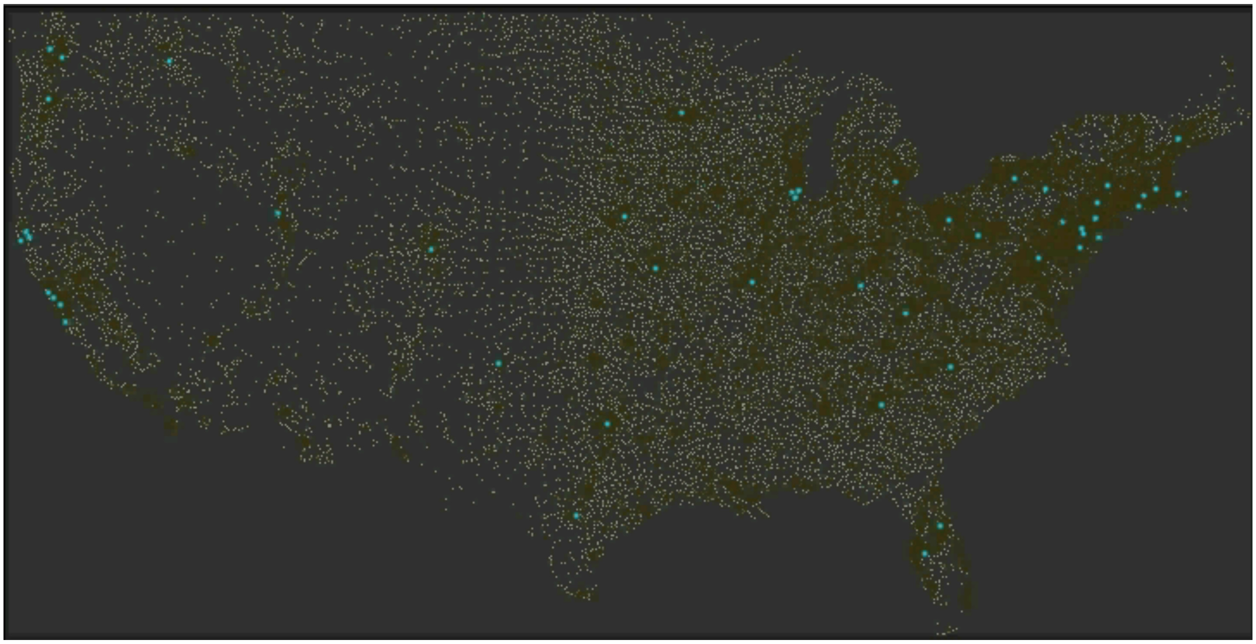
This initial design received negative feedback on a couple of levels. First, users were not clear as to how regions were allocated. Even after knowing that regions were split based off of the first zip code number, the response was that these distinction seemed arbitrary. This became rather apparent as well for New Jersey, which is included in the zip codes beginning with zero but is unattached to that region.

Second, coloring of zip code dots rather than regions themselves was somewhat confusing to our users. Due to the scarcity of zip codes in certain regions, such as the mid-west, those regions were difficult to see. We suggested fixing this problem by coloring regions uniformly, but ended up scrapping regions altogether.

Lastly, users were unsure if the dots represented individual Diggs in those regions. Since this seemed like a more natural representation of the dots, we now use dots to represent individual Diggs.

3.1.2 Progression of a Story

Our initial design also included creating the progression of Diggs for an individual Digg story. All zip codes are displayed and incoming Diggs appear on top of zip codes with the corresponding category color used for the Digg dot. While we still intend on creating the Digg progression for an individual story, we fused this idea with our previous initial design for the map to create a similar effect for Diggs in general.

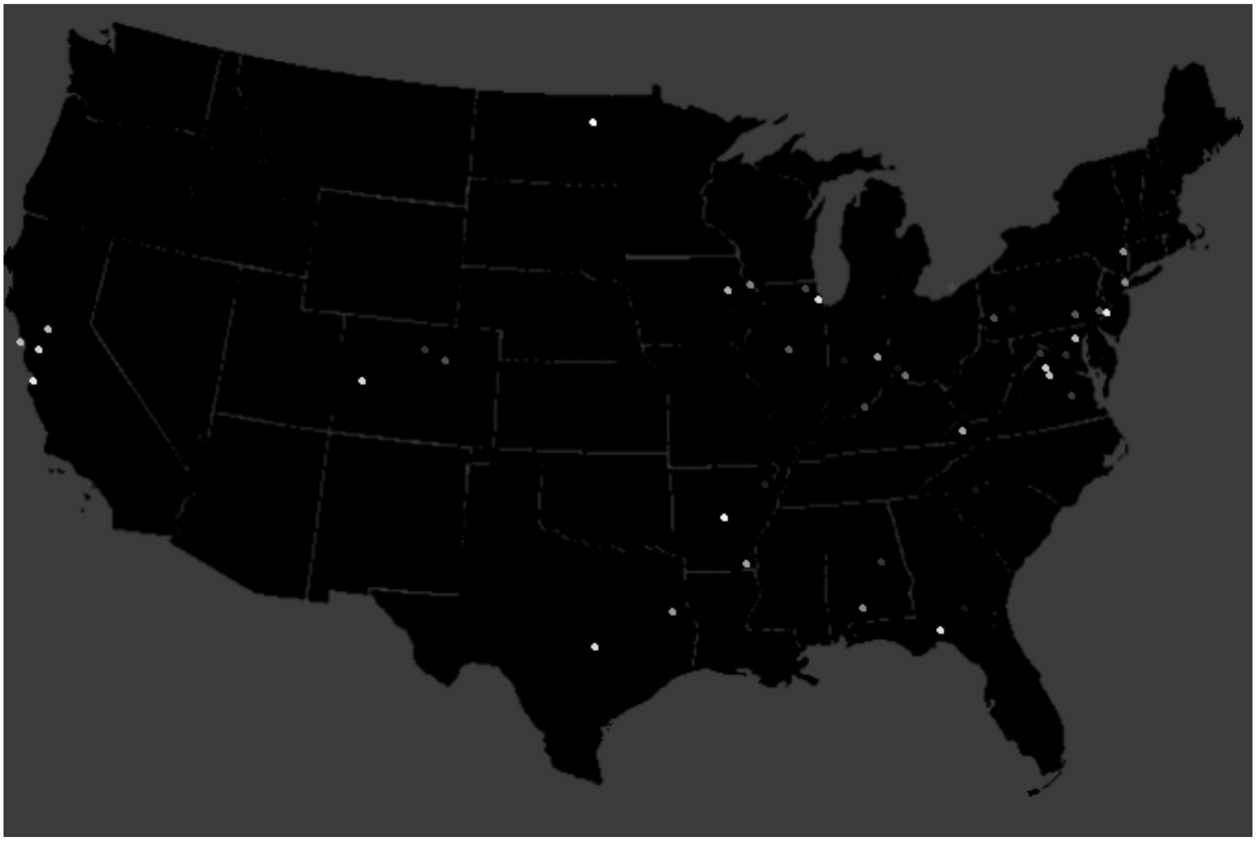


Initial progression design: individual Digg dots overlaid on top of zip code dots.

The response we received for this visualization was also problematic. First, the dots themselves were reported as being too small which made resulting progression less appealing to watch and individual Diggs were hard to see. Second, the activity dots placed on top of zip code dots made Diggs hard to see and made the visualization seem cluttered. To remedy this, we instead chose to show Digg activity on top of a simple black map of the United States. This made the dots themselves easier to see and still allowed for the geographic data to be shown. In addition, the inclusion of the map helped with sparser regions so that those regions did not look blank.

3.1.3 Final Map

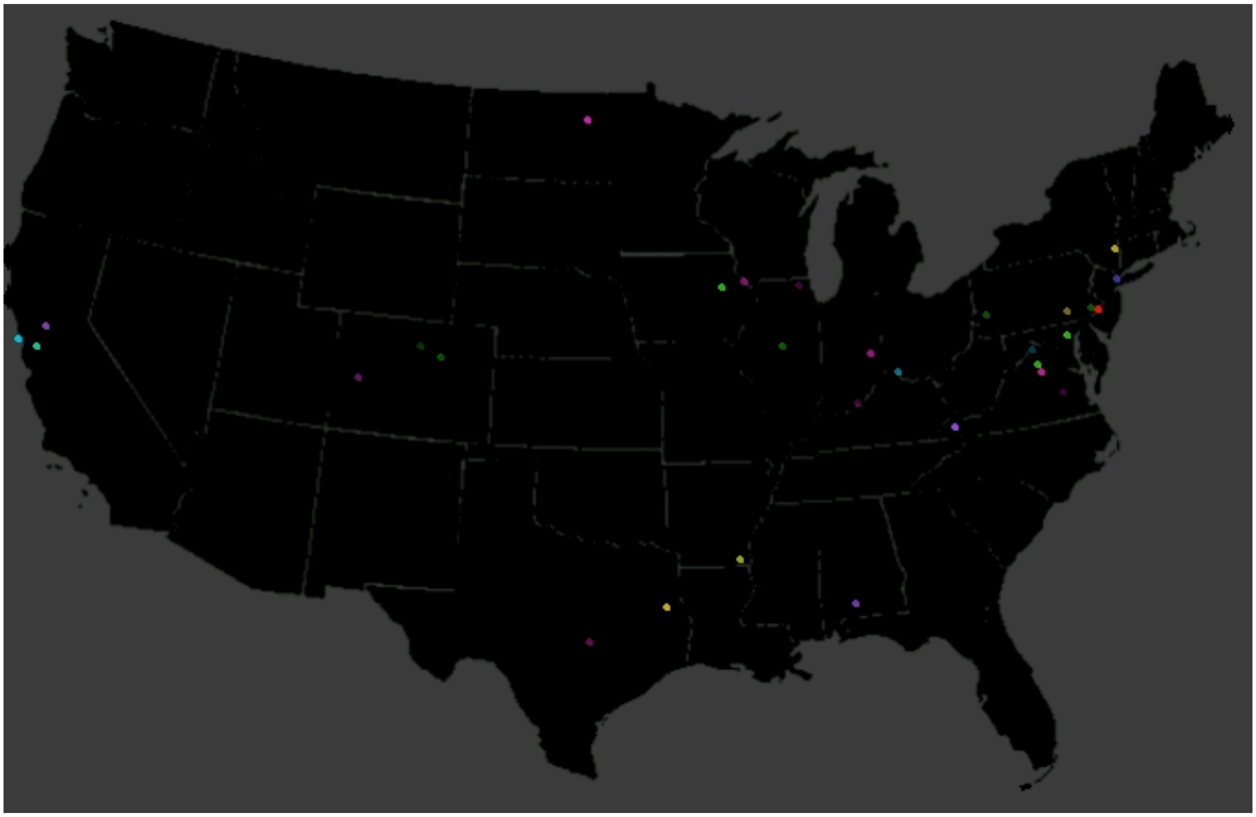
Our final map design took into account the feedback we got on the initial map design resulting in a much cleaner appearance. Digg activity appears on the map in its respective locations and fades out over time. Choosing a particular category displays only Diggs in that category. We originally thought of having the dots as somewhat transparent when appearing so that they could aggregate multiple values. Instead, for simplicity and visibility given the volume of activity, we have dots appear at full intensity and fade out slowly over time.



Final map design: dots representing individual Diggs overlaid over a black map of the United States.

3.1.4 Category Activity on Map

In addition to the category sparklines, we wanted to give the user the ability to see category activity as a function of location. For the total of all categories for Diggs, we were split between two variations: one with a single color and one with all the colors of Digg categories. The visualization with a single color was easier on the eyes and seemed to make more sense. However, having all category colors visible allowed for more pre-attentive processing and allowed users to detect which categories were more prevalent at a glance. As a result, we ended up creating both visualizations and plan to continue with experimenting with the two.

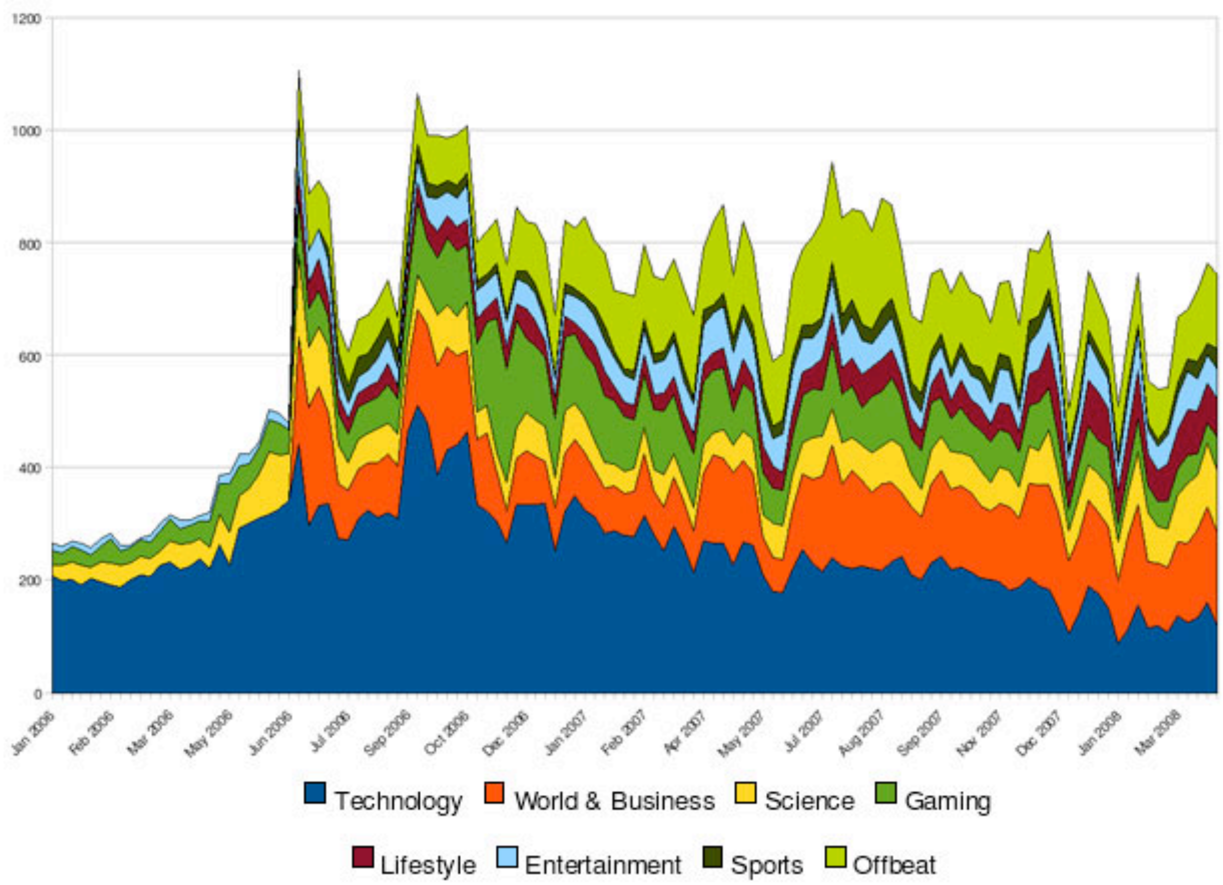


Final map design: includes coloring for the seven categories when viewing the total Diggs.

3.2 Category Activity Trends

3.2.1 Stacked Graphs

In order to visualize trends in activity by category, we initially considered creating a stack graph for each region similar to Martin Wattenberg's previous work, "The Baby Name Voyager"[5]. The initial design had users click on a region, which would bring up the stack graph on the side of the map. Our stack graph mockup actually looked very similar to the implementation seen on article "The Decline and Fall of Tech on Digg":

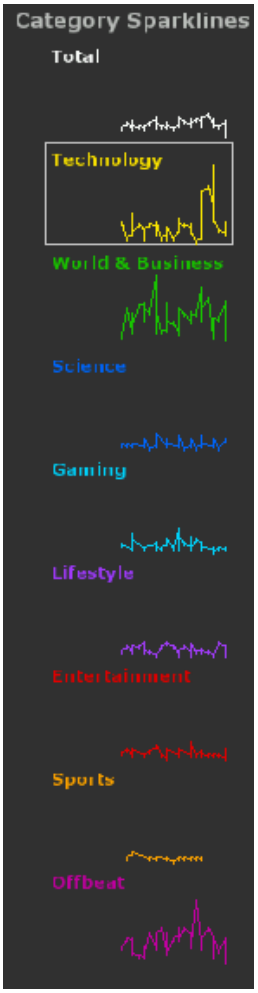


The Decline and Fall of Tech on Digg[6]

[5] http://www.bewitched.com/namevoyager.html
[6] http://www.readwriteweb.com/archives/digg_the_decline_and_fall_of_tech.php

3.2.2 Sparklines

The main critique of having a stack graph was the inability to see the actual category patterns for any category not placed at the bottom of the stack graph. Instead, we decided on having sparklines for each category so that users could see individual breakdowns as well as a total breakdown of Diggs.



Category Sparklines

For the sparklines, each category was given a different color that could as easily as possible be differentiated from the other colors such that viewing all categories on the map would still be differentiable.

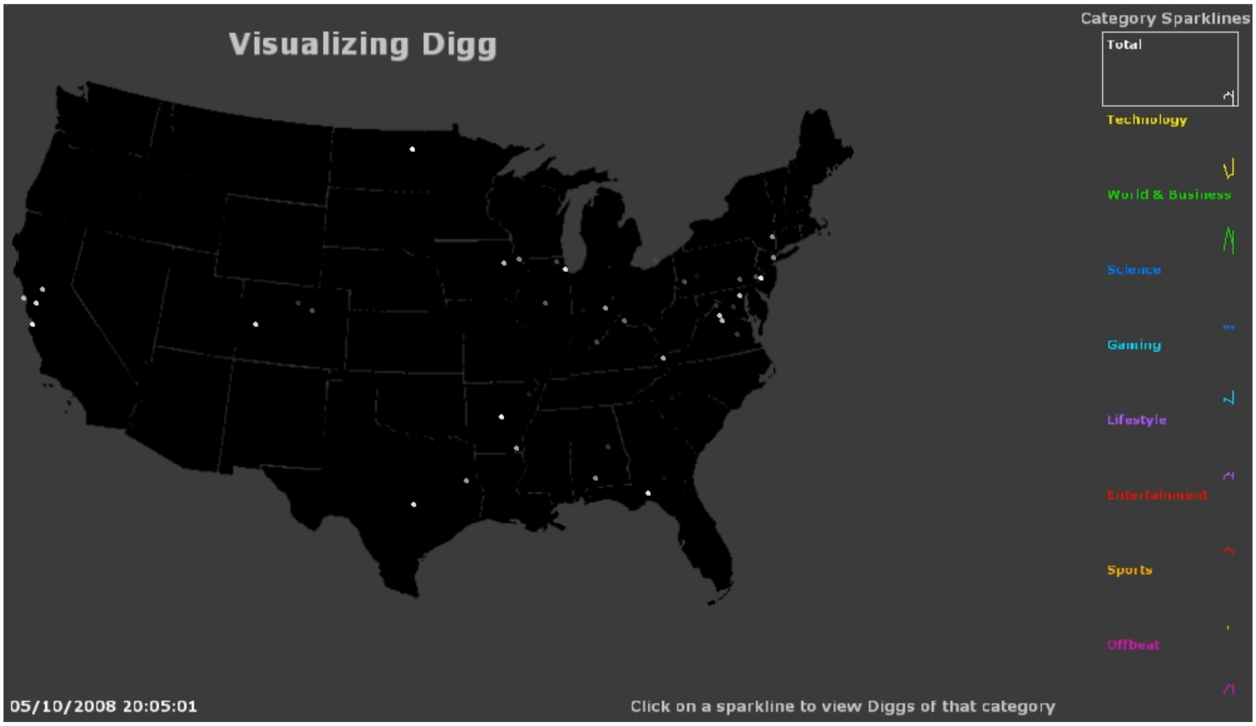
Initially, users would click on a region to bring up the graph. Since we removed both the regions and the single graph, we had to redesign the way we would link the map and the sparklines. After the suggestion for sparklines was mentioned, we created sparklines for each category that corresponded to all of the United States. Also, instead of clicking on the map to bring up the corresponding graph, users click on the graph to bring up the corresponding Diggs on the map.

To make the graphs look clickable, we included a highlighted box that appears when hovering over the region of a sparkline. A similar box will remain to indicate what category is currently selected. In addition to the visual cue when hovering, a message at the bottom of the visualization indicating the ability to click on the sparklines was included to provide users with a very clear idea of what they can do with the visualization, per the advice of one of our testers.

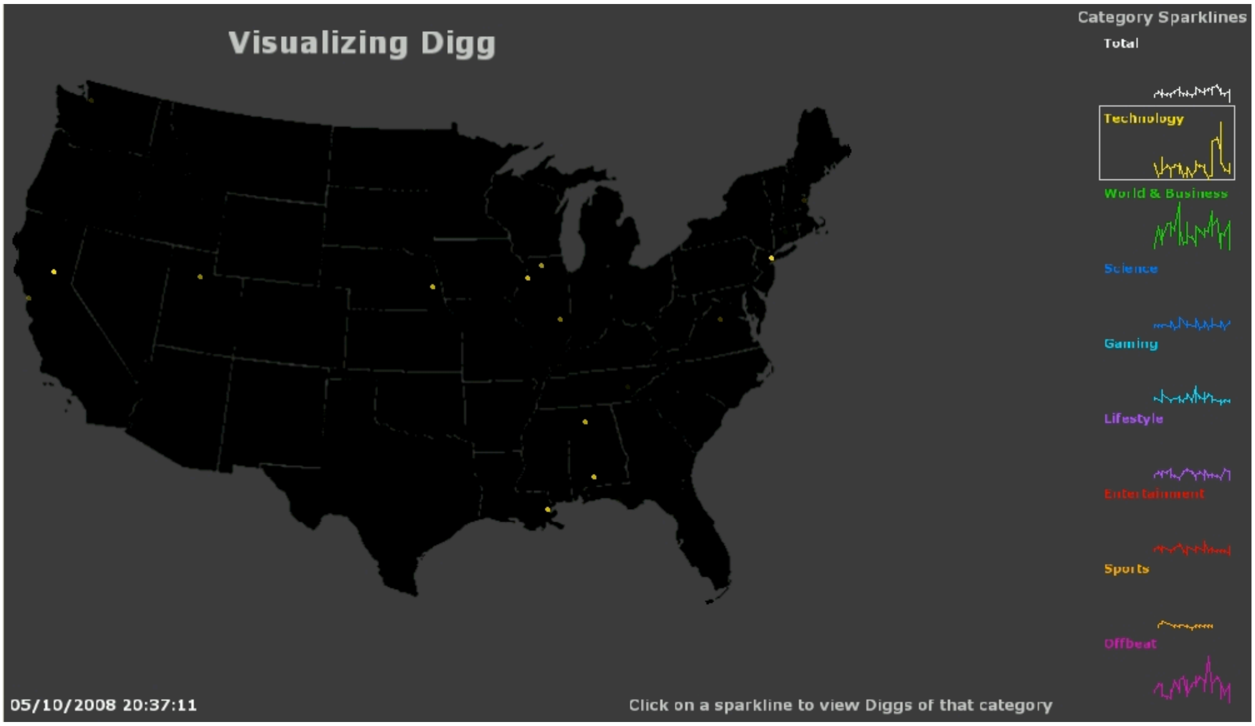
3.3 Timecode

As a final addition, to indicate the actual rate of Diggs, we included a timecode at the corner of the visualization. When the data source is live (vs. historic data), a message is displayed over the time state indicating that it is live.

3.4 Final Design

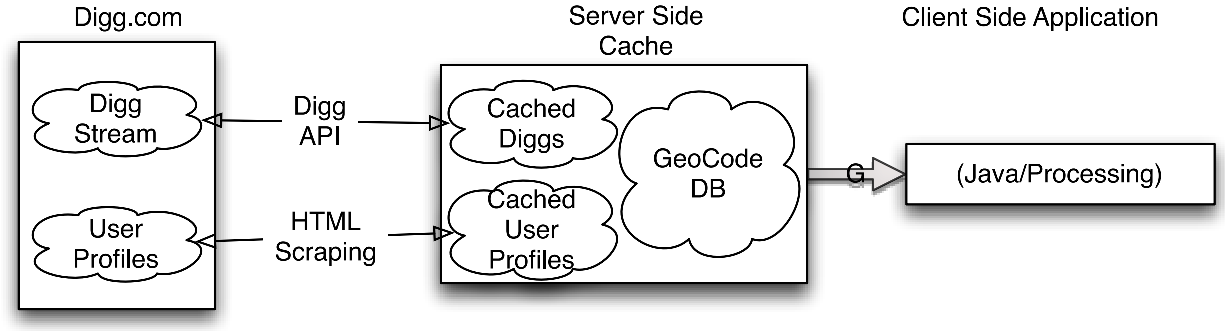


Final visualization: map and sparklines with time stamp.



Final visualization: final with technology selected.

4. ARCHITECTURE



Visualizing Digg Application Architecture

4.1 Data Source

To obtain real-time user Digg activity we use the API provided by the Digg.com.[7] However, since there were limits to the rate at which this data could be collected, and since some of the data necessary for our visualization, such as user profile information, was not available directly via the API, we employed a server-side component written in Python and MySQL to pre-process and cache the incoming Digg activity.

4.2 Workflow

At an interval of no more than once per minute, our server-side component connects to Digg.com via the API and downloads the latest activity since the last time it connected. For each incoming Digg, our program scrapes the corresponding user profile information from the Digg website and then parses that information for a country and city string [8]. If a string is found and that city is within the US, then it is matched against a known city:zip hash table.

For cities that have several zip codes (some have as many as 100), the user is assigned one at random from the list of zip codes for that city and then permanently stored in the system. Using random zip codes for the same city allows for jittering of repeated Diggs by different users in the same city on the map. Consistently using the same zip for a given user allows us to preserve the metaphor that a particular spot on the map represents activity from potentially a specific person.

[7] http://apidoc.digg.com/

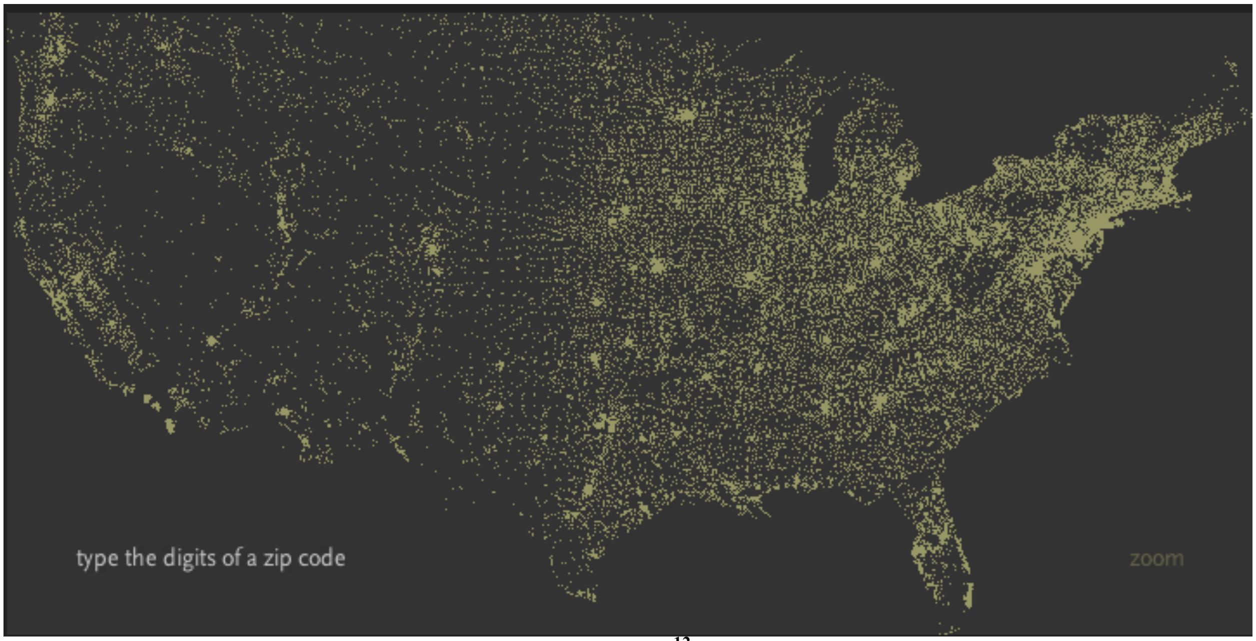
[8] User profile queries are also cached so that their profile page is only scraped once.

Users that are not from the United States are omitted as well as users that do not specify a valid city in the United States. Diggs for which a zipcode has been found are stored in our database and subsequently provided to the visualization front end in the following format [9]:

```
<timestamp> <story id> <category> <zipcode>
```

4.2 Visualization

In order to visualize the Digg activity stream, we use Java and Processing [10]. Some of the simple functionality, such as the mapping of zip codes to map locations and the graphing of points, were based off of the Zipdecode example in Ben Fry's Visualizing Data[11] book. Data structures from Java's library were also used for sorting and caching data. Our visualization utilizes Processing's 2D rendering for quick rendering of large amounts of data.



Zipdecode[12]

[9] A database was chosen in order to support multiple front end clients simultaneously with the ability to recall cached data from various time periods.

[10] http://processing.org/

[11] http://processing.org/learning/books/index.html#fry

[12] http://acg.media.mit.edu/people/fry/zipdecode/

5. IMPLEMENTATION

5.1 Analysis

The first major step was to manage the analysis of the data. This included the ability to determine where Diggs were coming from and the volume of those we were interested in. Because not all Digg users display their location, our main concern was whether or not there would be enough Diggs to display and whether or not the Diggs displayed would be representative of the Diggs overall.

Additionally, we found that the actual activity was extremely 'bursty' in nature. The rate of Digg activity varied substantially depending on the time of day and, as a result, visualizing this information in real time was uneventful due to periods of low activity.

5.2 Functionality

After processing the data, several parts of the visualization required individual attention. We started out with the map and the display of Diggs. Things to consider for this section included the size of the dots for the Diggs, the rate at which the Diggs should appear, and the rate at which the Diggs should fade out. We found that depending on the rate of Diggs, different values for the rate at which dots faded out were appropriate. I.E. If the fadeout was too large, then the map would look too overwhelming during high-volume periods.

5.3 Sparklines

After handling the map, we focused on the sparklines. Some issues with the sparklines were the scales for the axes, the coloring for the categories, the continuity of the graphs, and how the graphs would progress. We iteratively tried different ranges for the sparkline scaling and finally settled on one that seemed to work well for all periods of the day. We eventually want to scale the sparklines dynamically based on volume in a future version.

5.4 User Interface

In addition to the displaying the sparkline, we had to figure out how a user would know to click on a sparkline in order to select a category. Initially, we had it so that the user would click on the actual line, which was actually somewhat difficult to do, especially as the graph moved. Instead, we allowed for the user to just click in the area where the sparkline appears, a box that includes the title of the graph and the invisible axes. This also made it easier to show a highlight of selected sparklines as a box was just drawn around the boundaries of the graph.

5.5 Visual Aesthetics

A major portion of our work was tweaking the appearance of the visualization in order to make it aesthetically pleasing. Since we were going for an ambient display, we felt one of the key factors to make our visualization successful was to make it enjoyable to watch. This included changing the drop off rate of Diggs, changing the rate at which Diggs appear, changing the time frame of Diggs prior to the present, and changing the color corresponding to the categories.

6. RESULTS

Originally, our goal was to show live Digg activity across the United States in an attempt to demonstrate a sense of *presence* through our visualization. After viewing the actual progression of Diggs however, we came to the conclusion that watching live Digg activity was not as exciting as we had hoped. This was partly due to our overestimation of the rate at which Digg's actually occur and partly due to some limitations resulting from the way we're forced to interface with the Digg API. As a result, we receive qualified Diggs at about once every two seconds on average, resulting in a relatively boring live visualization.

Instead, we modified the visualization to playback incoming Diggs at an accelerated rate, providing a much more captivating visualization in our opinion. We'd like to continue to experiment with this in the future and explore ways to visualize 'live activity' as well as provide interactive features such as the ability to select time periods or individual stories to display.