Sh\*t HN Says: A Week of Hacker News

By: Brian Connolly and Peregrine Badger

# Project Proposal:

Tentative project title: Hacker News visualized.

## Name, email, programming comfortability of each member of your group

Peregrine Badger, pbadger@college.harvard.edu, very comfortable

Brian Connolly, connolly@college.harvard.edu, very comfortable

## Research questions and hypotheses:

Sample HN questions, followed by hypothesis:

1. “Whats the average number of upvotes a post gets?” 20

2. “How long do posts stay in ‘top posts’?” a day

3. “What pages are the posts pointing to?” nytimes, github, nature

4. “What programming languages get the most coverage?” javascript

5. “What topics are on HN right now?”

6. “What times of day is Hacker News most active? (in terms of number of posts / time interval)

## Motivation

Hacker news is the primary news source that Peregrine follows, and is pretty awesome for getting relevant tech news that is up to date. When a Rails insertion exploit was discovered, it was on HN before any Rails based site (as far as I know) got hacked. Still, its layout is just a bunch of text, so it would be nice to visualize the topics in an intuitive way.

Data

What data will you use to construct your visualization? How is it obtained? How is it relevant to your research questions? If appropriate, provide a link to your data source.

The data source will be both:

1. Using the Hacker news api. The API returns a JSON object with the title, the post’s url, a post ID, the number of upvotes, the number of comments, and a time and date.

2. Scraping http://news.ycombinator.com/ My questions are about Hacker news, so this data is exacty what I need. We’ll use the same python libraries we used for pset 2, namely Pattern and Mechanize.

Visualization

How will you display your data? Provide some general ideas that you have for the visualization, broken down into two tiers of priority: those ideas that your project will aim to implement and those ideas that, given time, would ideally be implemented. You will be asked to sketch some designs in the official project proposal submitted with HW5.

1. Probable visualizations (sketches shown at end):

- Bubble chart : we will have a chart of bubbles for each post on HN whose size is based on number of upvotes of the post. You will be able to click on each bubble to get more information about the post, such as URL and top keywords for the post, possibly in a word cloud on the side. The color of the circles may be based off age and you could filter the bubbles by most popular or by age.

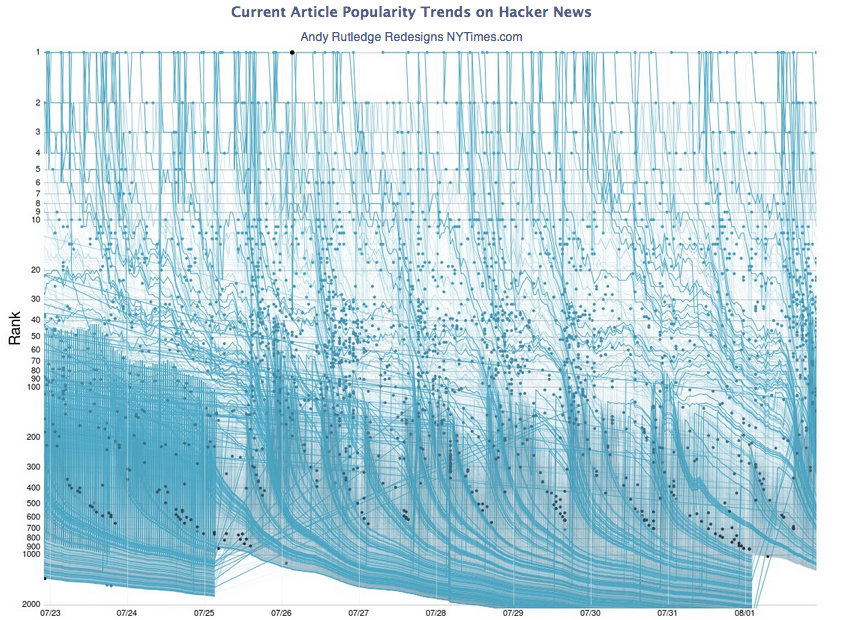
- Line graph : the line graph will primarily look at the frequency of posts over a given day. The hope is to be able to get a glimpse of when HN is most active. There will be 7 lines layered over each other, one for each day of the week. You will be able to compare activity on weekends vs. weekdays. When you click on a certain time of day, you will see the list of posts associated with that time interval.

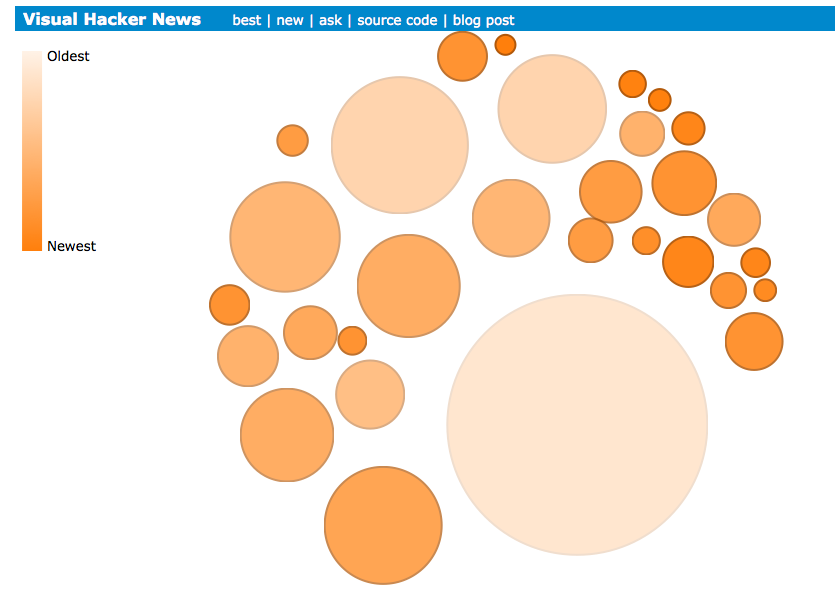
- Word cloud : In this visualization, a word cloud would initially fill the screen, but when a word was clicked, a dashboard would slide out that would have the posts with that keyword, ranked by upvotes. The word cloud would be generated using text scraped from the links on hacker news.

2. Ideal visualizations:

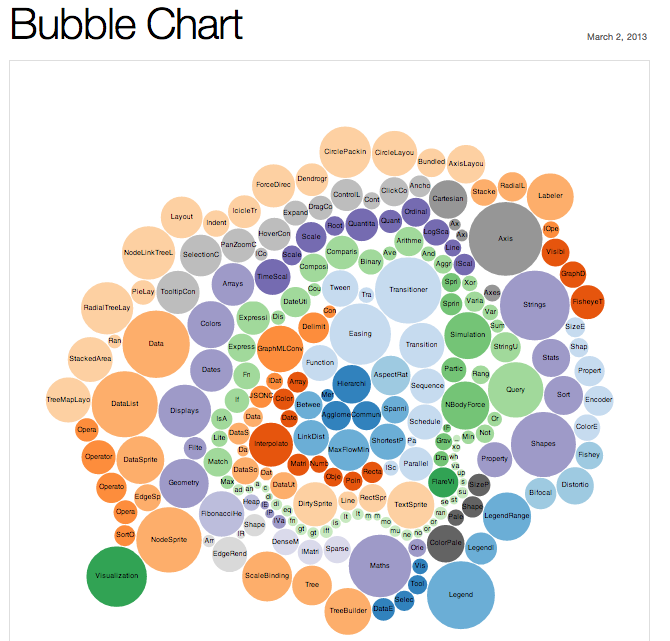
-Line graph: ideally with the line graph visualization above we will also put a word cloud for each time interval to see if there are any popular keywords for a given time of day. For example, when the pope was announced on Wednesday, there could have been a spike in the usage of the word “pope” and we would be able to see that.

Some sample visualizations that have been done on HN are:



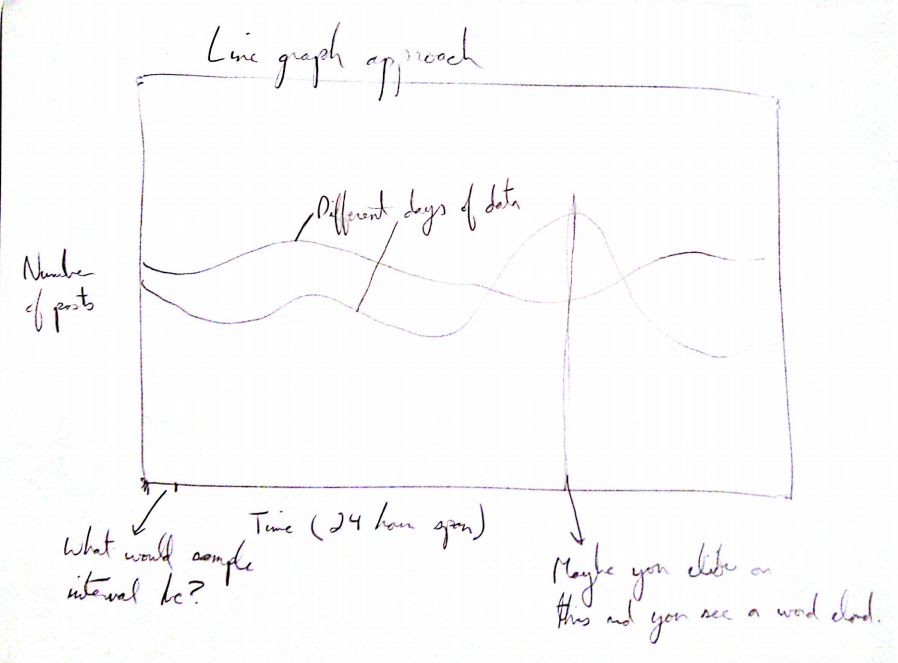


But these are very basic, and don’t allow for any higher level ordering of articles. My visualization would be more like:

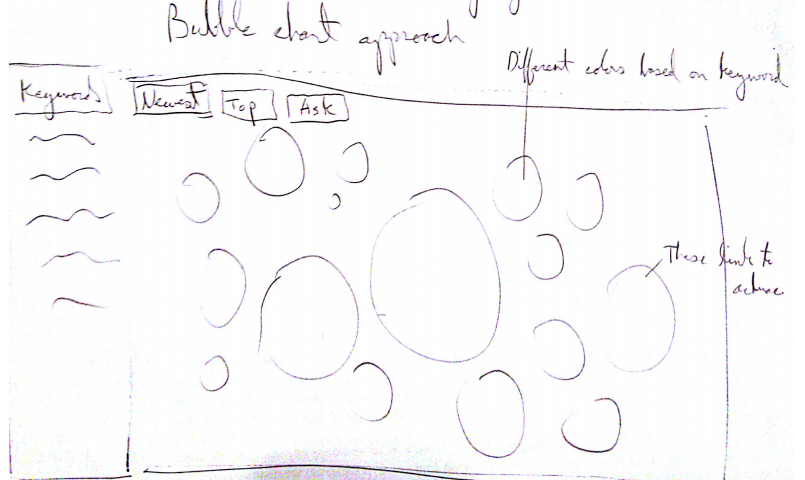


All of the colors would represent different topics, and hovering over a bubble would have the same effect as the “visual hacker news” above. This is probably the “first tier” of attempt.

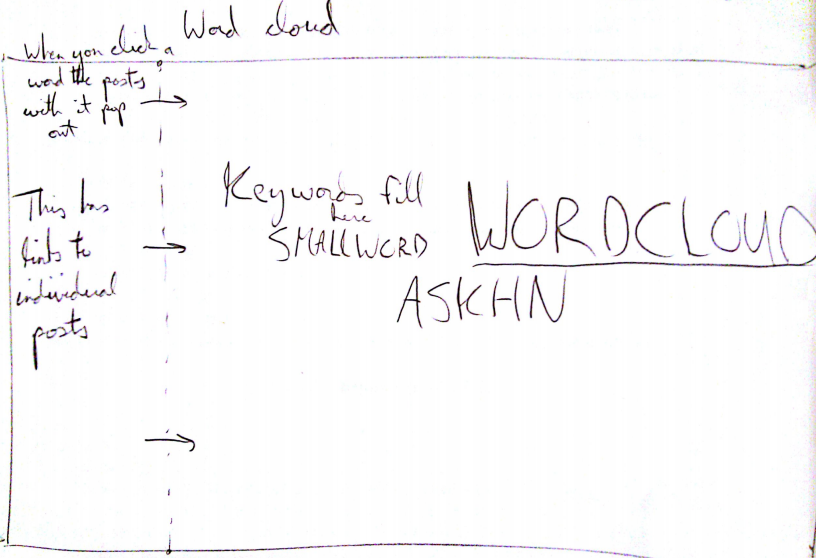
## Sketches



In this line graph, the height would represent the number of posts in some time interval, and there could be a different line for each day of the week.



In this bubble chart, clickable bubbles get colored based on which keyword they have. The bubbles could link to the url of the post.



In this visualization, a word cloud would initially fill the screen, but when a word was clicked, a dashboard would slide out that would have the posts with that keyword, ranked by upvotes.

## Tentative final plan:

Main view: Line graph with weeks worth of data, split by hour.

When hour is clicked, bar graph fades in with every post from that hour. (Heights of bars are number of points post has).

On the side, another bar graph shows up that has the 10 most common words. When one of these is clicked, all the article bars with that keyword light up (This is linking/brushing).

Data collection: every 12 hours pull last 12 hours of articles, then visit each link and process the articles for keywords (use wordcloud js as example).

# Process

Along the way our final plan changed slightly. We decided to ditch bubble graphs all together because they do not encode that much information given their scope.

Our first visualization is the line graph. We went with nice distinct colors for the different lines to promote readability. Originally we were going to put text labels for each day directly next to the end of each line, but since the lines were relatively close together so we went with a legend next to the graph.

When you click on a day in the line graph, the other side of our visualization comes to life. We decided to go with a scatterplot based on the number of points you receive. We believe the scatterplot was a good choice because it encodes a lot of data and it is easy to pick out the most popular posts given a specific hour.