Not Safe Anymore (NSA)

Background and Motivation

Web Security is an on-going concern for all organizations and individuals. It is also a time and resource consuming exercise that needs to be balanced against other priorities. Given the recent online attacks on big banks[1] and retailers[2], we are interested in visualizing the attributes of such attacks - which countries do they target the most, what times of the day are popular for such attacks and what networks (ISPs) are they carried out on.

Project Objectives

As an individual interested in using the Internet for online shopping and financial transactions, I want to understand the risks associated with my actions when I am online. By looking at various visualizations of the data we hope to show both geographic and cyclical trends, and help discover unexpected patterns.

By answering the following questions through the visualization, we aim to provide our users with a better understanding of their risks while online:

- What geographic and cyclical patterns are present in the data?
- What time of day are the risks lowest?
- Which ISP's provide the best security?

Data

This dataset is a sample taken from the firewalls running on Akamai's[3] global network[4] of servers that serve web content all over the world. These firewalls log and report each intrusion attempt to Akamai's monitoring systems. This dataset was scrubbed of sensitive information by dropping dimensions that identify customers and only the keeping the dimensions needed for this project. Although Akamai collects data points on each firewall intrusion, given the volume of data and our visualization goals, we have decided to aggregate the samples on a per 15 minute basis. Finally, we are planning use data collected for a 24 hour time period for this project.

Our data structure has the following variables:

Field	Data Type
`timestamp`	Quantitative (interval)

[1] Cyber attacks against banks more severe than most realize, http://www.reuters.com/article/2013/05/18/us-cyber-summit-banks-idUSBRE94G0ZP20130518
[2] Target cyber breach hits 40 million payment cards at holiday peak, http://www.reuters.com/article/2013/12/19/us-target-breach-idUSBRE9BH1GX20131219
[3] Akamai, http://www.akamai.com/
[4] Akamai's global network of ~ 137k servers, http://en.wikipedia.org/wiki/Akamai_Technologies#The_Akamai_Network:_Edge_Platform

`country`	Categorical
`state`	Categorical
`city`	Categorical
`lat, lng`	Quantitative (ratio)
`number_of_attacks`	Quantitative (ratio)
`ISP`	Categorical

Data Processing

Because we are using data generated and logged by web servers, they are already in a machine processable format. In addition, the scrubbing of personal information is occuring before we receive the data. We may need to perform a minimal amount of data cleaning and transformation, such as converting timestamps or formatting ISP names.

Visualization

There are a number of visualizations that we expect will be useful for exploring this data.

Overall Filters

Filters should be displayed at the page level, and affect all relevant visualizations. These could include:

- Scrubber for time of day
- Check boxes to filter by geography
- Check boxes to filter ISP's (eg. Top 9 and Other)

Individual Visualizations

In addition to the global page elements, there are other visualizations that we feel might be useful:

- Maps:
    - thematic maps with bubbles overlay
        - Bubble area indicates number of attacks for that region
        - Color of bubble for ISP
    - map of continents with polar plots for events at each time of day
- Timeseries:
    - spiral plot of attacks over time for top 10 countries
    - stacked area chart by top X countries (rest could fall in "others" category) showing attacks per hour
- Hierarchical:
    - icicle plot showing total number of attacks broken by country, state and city
- Tabular:

    - top 10 cities by number of attacks based on the currently selected filter (see Overall Filters)

Must-Have Features

The visualizations must have:

- a spatial visualization to depict web attacks all over the world
- a hierarchical breakout of attacks by geography
- a time series visualization showing attacks per unit of time
- the ability to filter the visualization by time (by the hour)

Optional Features

These features would be nice to have:

- suitable animation effects for transitions
- the ability to filter the visualization by country
- the ability to filter the visualization by ISP

Project Schedule

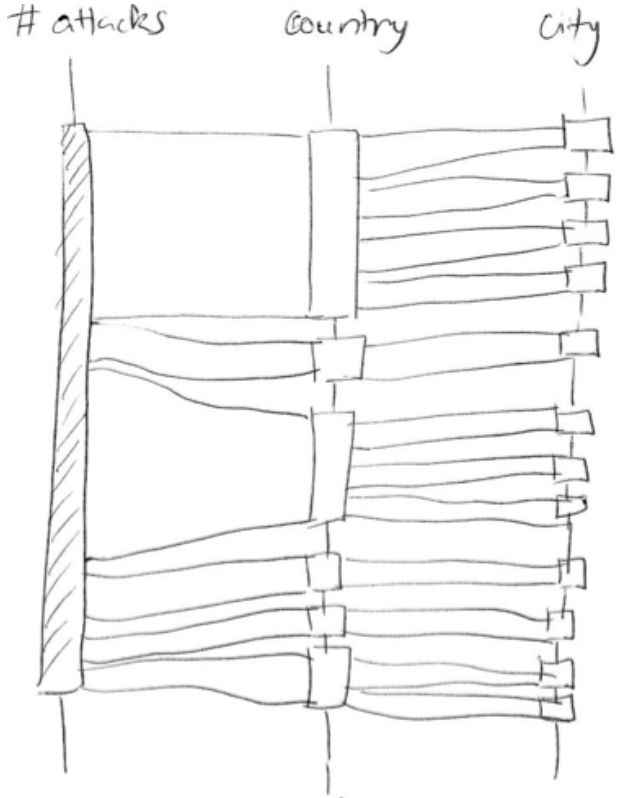
In general, the work will be split into visualization elements that each person can develop independently and then display in a defined bounding box. This allows for parallelization of the effort. There will be weekly check-ins to ensure that progress is on-track. The schedule, including the required milestones, will be as follows:

- Thursday, March 13: Project proposal due (part of Homework 3)
    - March 20 - identify initial set of visualizations
        - Chris: research polar and hierarchical visualizations
        - Shreyas: research spatial and time series visualizations, overall design for the page
    - March 27 - prototype a subset of these visualizations, validate visualization choices made so far
        - Chris: implement draft of polar and hierarchical visualization
        - Shreyas: implement draft of spatial and time series visualization
    - April 3 - iterate and update initial set of visualizations based on the nature of the data and findings from prototyping
        - Both: finalize in-progress visualization
- Thursday, April 10: Functional project prototype due
- Week of April 14: Project review with the TFs
    - April 17 - implement remaining visualizations and incorporate feedback from the project review
        - Both: update and integrate visualizations, including links between visualizations
    - April 24 - continue with the implementation
        - Both: finishing touches and clean-up; documentation

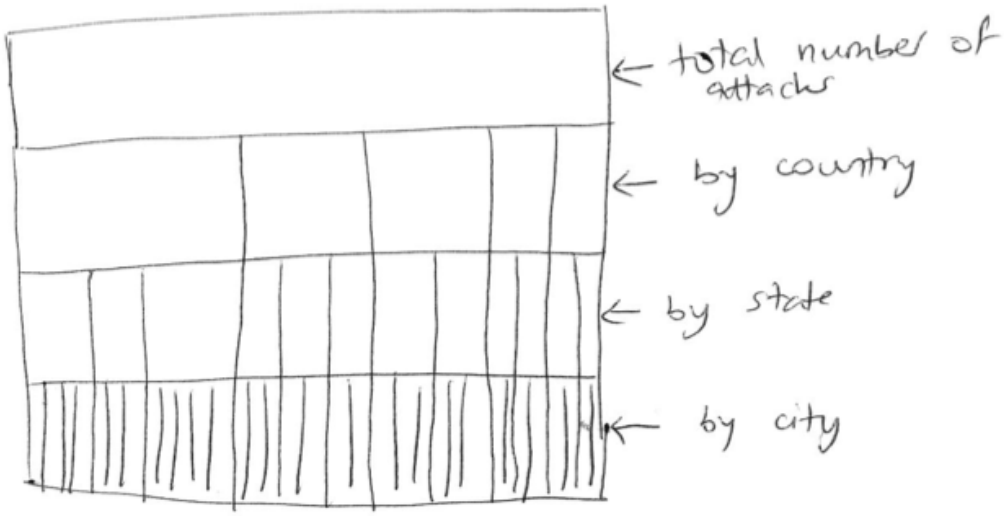
- Thursday, May 1: Projects due (including screencast)
- Thursday, May 8: Best project presentations and prizes

Appendix

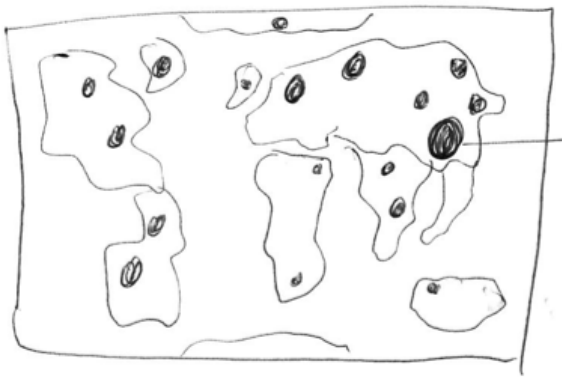
Mockups



sankey → number of attacks by country by city

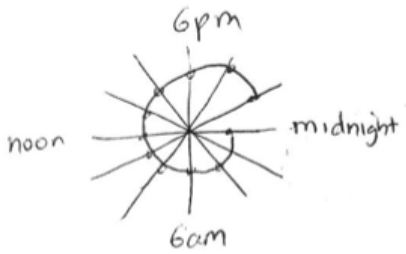


icicle plot



area ~~size~~ of bubble = number of attacks

color = by ISP



Spiral plot of attacks over time for A country