Carnegie Mellon University
CyLab

Student Seminar Series

Topics of Controversy: An Empirical Analysis of Web Censorship Lists

Friday, 28th April 2017
12:00 – 1:00pm
CIC 2201

Speaker : Zachary Weinberg

Zachary Weinberg was born in Los Angeles, CA in 1978, but escaped at the earliest opportunity. He has spent the years since doing, variously, chemistry, C compiler maintenance, cognitive linguistics, distributed version control system development, Web browser development, and Web security, before a fateful internship at SRI in 2012 put him onto censorship circumvention and measurement.

Talk Abstract:

Studies of Internet censorship rely on an experimental technique called *probing*. From a client within each country under investigation, the experimenter attempts to access network resources that are suspected to be censored, and records what happens. The set of resources to be probed is a crucial, but often neglected, element of the experimental design.

We analyze the content and longevity of 758,191 webpages drawn from 22 different probe lists, of which 15 are alleged to be actual blacklists of censored webpages in particular countries, three were compiled using *a priori* criteria for selecting pages with an elevated chance of being censored, and four are controls. We find that the lists have very little overlap in terms of specific pages. Mechanically assigning a topic to each page, however, reveals common themes, and suggests that hand-curated probe lists may be neglecting certain frequently-censored topics. We also find that pages on controversial topics tend to have much shorter lifetimes than pages on uncontroversial topics. Hence, probe lists need to be continuously updated to be useful.

To carry out this analysis, we have developed automated infrastructure for collecting snapshots of webpages, weeding out irrelevant material (e.g. site "boilerplate" and parked domains), translating text, assigning topics, and detecting topic changes. The system scales to hundreds of thousands of pages collected.

This is a practice talk PETS 2017: https://petsymposium.org/2017/

Please direct any questions regarding this seminar to Soo-Jin Moon (soojinm@andrew.cmu.edu) or Mahmood Sharif (mahmoods@cmu.edu)