This document include both some high-level, open ended project choices and some slightly more specific project descriptions. For each of the projects, I ask that you do not share datasets and developed tools publically until we potentially try to publish a research article using these tools and datasets (at which point things hopefully have been sanity checked, is polished, and the tools/datasets can help improve the odds that such paper is accepted and published). Also, please discuss and update me on your ideas and progress so that we together can try to make the most of the class projects.

1) Descriptive title: Data collection and analysis of phishing domains
In this project you will try to identify as many complementary data sources (e.g., public records, CT logs, traffic traces, routing tables based on prefixes, geo-data, etc.) as possible that provide information about known phishing domains (e.g., https://www.phishtank.com), and then perform a preliminary data collection and analysis that help address a number of research questions. As part of this project you will help implement a data collection framework, perform large-scale data collection using this framework, perform careful pre-processing of the dataset so to create a cleaned up dataset, and ideally also perform a preliminary analysis in which we answer a number of research questions related to phishing domains. The goals is that the datasets collected with the tool can be used to help answer some example research questions. This project require good programming skills. Familiarity with using web APIs is also recommended. Please discuss your research questions and data collection ideas with Niklas. While some ideas exists here, it would be good to see if there are other interesting datasets that can be used (e.g., as shared in prior measurement conferences and via other related/non-related papers).

2) Descriptive title: The social networks of the gaming communities
Outside the games, users may socialize in numerous ways, including by commenting on the gamecasts (i.e., records of games) and chatting with their friends through various online resources. For example, some popular online game communities provide an interactive gamecast sharing service, wherein the creators promote their gamecasts through live streaming with on-air explanations (in audio and text format) of their game styles. In this project you will develop a measurement methodology, collect data, and present a preliminary analysis of the social networks formed in one or more such communities. Of special interest are the social interactions (which in some cases can express the strength in user relationships, for example). Also, do you find heavy tailed relationships or other interesting characteristics?

3) Descriptive title: IPv6 routes (and IPv6 adoption in general)
Tasks: Use trace routes to compare and contrast the routes taken when using IPv6 compared to when using IPv4. Are there any differences in the routes? Impact of the current state of IPv6 adoption? Develop a methodology, collect measurement data (either existing or such you collect yourself through measurements), and present a preliminary analysis.

4) Descriptive title: Measuring the third-party [tracking] usage and effectiveness of adblockers
Tasks: Create a measurement methodology and perform large-scale measurements to evaluate the third-party tracking usage and the effectiveness of adblockers. Could leverage and extend some existing code from our LCN 2016 paper (https://www.ida.liu.se/~nikca89/papers/lcn16a.pdf), for

example. Develop a methodology, collect measurement data (either existing or such you collect yourself through measurements), and present a preliminary analysis.

5) Descriptive title: Spreading of "Fake" News

Description: Fake news and other misinformation distributed through social media currently threatens and undermines our entire society. However, measuring and modeling how fake news and biased content is shared is non-trivial. In this project, you will develop a methodology (likely involving both manual and semi-automated classification of articles and the tweets/retweets related to these articles (about 3-4000 articles identified during spring 2018 and spring 2017; provided by Niklas), including the number of followers of tweeters and re-tweeters, and all other information that may help model and understand the information propagation on twitter. You should try to collect as much of the information propagation related to these articles as possible, as well as other meta information about the articles (e.g., if possible manual classification of article type/category). Try to be precise in your definition of what is "fake news" (e.g., fact related, biased, etc.) and try to develop the methodology and tools so that we can use it/them, at larger scale, after the course. A solid initial dataset together with some preliminary analysis would also be expected. Note that the methodology can involve crowd sourcing or other means, but would in such a case need to capture some measure of confidence in the results.

6) Descriptive title: Characterizing mobile web adaption and any correlation with third-party [tracking] services

Tasks: The project would follow a combination of methodologies to collect measurement data for popular websites when using different user agents (mobile vs non-mobile, for example). The goals would be to (i) carefully implement a measurement methodology, (ii) perform data collection, and (iii) present a preliminary analysis.

Requirements: Access to a machine (e.g., laptop) with root access is helpful, such as to allow various measurement tools (e.g., bro) to be used and user-agent headers to be modified/faked, for example.

On reading list: M. Butkiewicz, H. Madhyastha, V. Sekar, "Understanding Website Complexity: Measurements, Metrics, and Implications", Proc. IMC, 2011, pp. 313--328.

7) Descriptive title: Comparing tails: Identification and collection of long tail datasets

Goal: Create a repository of uniformly formatted datasets that with potential long-tail properties. This would include a combination of identification of public distribution-related datasets (with different forms of network related data), but could also include collection or creation of additional datasets (e.g., by extracting the distribution data from other sources). All files should follow a uniform row-column format, and a high level analysis that compare and contrast the datasets should be performed.

A quick read: A. Mahanti, N. Carlsson, A. Mahanti, M. Arlitt, and C. Williamson, "A Tale of the Tails: Power-laws in Internet Measurements", IEEE Network, Vol. 27, No. 1, Jan/Feb. 2013, pp. 59--64.

8) Descriptive title: Beyond existing measurement work ...

Tasks: Pick a paper or two from IMC (e.g., one that we looked at in class), identify some related issues/problems, and try to collect your own large-scale dataset that help answer these related questions. For these open ended projects it is particularly important that you discuss your research questions and data collection ideas with Niklas.