TDTS21 projects vt 2025

This document includes a mix of open-ended projects and projects with specific ideas (even if only high-level descriptions here).

Data sharing statement for all projects: The datasets, tools, and analysis should not be shared publicly until we potentially publish a research article using these tools and datasets. (If needed, at such a time we hopefully have been able to sanity check, polish, and improve the tools/datasets. Postponing the release to such time also significantly improve the odds that a paper is published.) However, to enable continuous research and potential publication, already at the end of the term, all code, data, text, and results must be shared with your supervisor. 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 (regardless if it has potential to be published or not).

1) Descriptive title: Build a Selenium-based crawled to look for some interesting aspect
In class, we observed some examples where a Selenium crawler was used to evaluate some properties of the websites listed on a top list (e.g., Tranco). In this project, you would build or modify an existing tool to collect statistics regarding some aspects of interest to you. For example, could consider the use of single-sign-on, following up on this work:
https://www.ida.liu.se/~nikca89/papers/networking22b.pdf
or some other topic of interest.

2) Descriptive title: Characterize the security properties of the web
In this project you will first identify different free (but relatively reliable) tools to evaluate the security properties of a website. Second, you are expected to use these tools to evaluate the security properties of a large set of websites (e.g., based on the Tranco list) and provide the output in well-structured data files (e.g., csv). Finally, you should provide a preliminary analysis in which you compare and contrast both website classes against each other as well as the results of the different tools (e.g., which tools seem most reliable with regards to what aspects). Ideally, we want to find some good, useful tools to capture the current state, with the potential to use these tools and comparison to gain some new insights.

3) 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) and the amount of additional network traffic generated around a live event (both in parallel and afterwards). Also, do you find heavy tailed relationships or other interesting characteristics?

4) Descriptive title: BGP interceptions in the wild

Identify potential BGP interceptions that have taken place recently and characterize them (accidentally or intentionally) using similar methods as in our 2013 paper https://www.ida.liu.se/~nikca89/papers/pam13.pdf

5) Descriptive title: Does being a stereotype help becoming targeted or is it enough visiting a page?

Build a Selenium-based data collection framework that generate web sessions based on different user profiles and compare what level of personalized ads (i.e., ads that match the user profile) that they see on the pages that they visit compared to a plane user (that have no history) going to the same website. Here, you will need to generate user traffic for some time and then collect how the ads presented to these users change over time. The goals here is to design a methodology and tool to (in a good way) measure the level of personalization we obtained just by ad campaigns targeting users of certain pages compared to ads that actually target individuals regardless of what website they visit. The use of adblockers and/or cookie settings can be alternative dimensions that can be considered here. Some related works from the group here include:
https://www.ida.liu.se/~nikca89/papers/wpes21.pdf
https://www.ida.liu.se/~nikca89/papers/mascots22.pdf

6) Descriptive title: Follow the Money of Crypto Currency Spam

The last few years, I have had various projects in which we have looked at how much money that different bitcoin addresses included in spams pointing to Bitcoin addresses (e.g., threat emails) attract. One outcome of these projects is a recently published paper:
https://www.ida.liu.se/~nikca89/papers/pam24a.pdf
The idea with this project is to try to perform a more thorough analysis of the money flow. While there exists some code, the idea would be to update the multi-step analysis and perform such analysis on various subsets of addresses (e.g., to better understand mixers, often used to disguise money flows). Here, you will need to work with several public APIs and need to put yourself into prior code and dataset.

7) Descriptive title: Current IPv6 deployment

Use measurements to capture the adoption of IPv6 from different perspective; e.g., using different forms of public data (e.g., DNS records, RIR records, RouteViews data, other BGP data) and potential experiments (e.g., traceroutes). For example, consider some of the methods described by J. Czyz et al., "Measuring ipv6 adoption." ACM SIGCOMM Computer Communication Review 44.4 (2015): 87-98.

8) Descriptive title: Certification Transparency

Download and analyze a popular CT log. For example, try to identify all certificates that have been registered in the past X weeks and come up with a methodology to classify whether each certificate is for a domain that (i) is likely to be used for phishing attacks or drive-by-downloads (e.g., use similarity

metrics of domain names to flag them as suspicious, for example, but I suggest that you do not visit those domains yourselves), (ii) have domain names that are covid or health related, for example, and (iii) some other classes that may be worth an extra look at. Also, please extract information about keys, hashes etc. (similar as in J. Gustafsson, et al., "A First Look at the CT Landscape: Certificate Transparency Logs in Practice", Proc. PAM 2017) for all certificates. Using this classified dataset, we can then look at trends in different classes and compare it with the baseline of "all" certificates registered.

9) Descriptive title: Recreating existing measurement work on a topic of interest ...
Pick a measurement paper of interest (e.g., perhaps from the set of papers we have seen already) and try to collect your own large-scale dataset and try to (1) validate the results (or part of the results) presented in that work and/or (2) try to extend the work in some direction. As with all projects, it is important that you discuss your research questions and data collection ideas with Niklas.

10) Descriptive title: Beyond existing measurement work on a topic of interest ...
Pick a topic of interest to your group (e.g., perhaps based on a paper or two from IMC, and identify some related issues/problems) and try to collect your own large-scale dataset that help answer some interesting/important questions. As with all projects, it is important that you discuss your research questions and data collection ideas with Niklas.

11) Descriptive title: Traffic analysis attacks and defenses
Removed text. Will share with Ethan, Somiya, Sheyda.

====