Data statement for the project ”*A Matter of Tech Security*”.

Data consists of transcripts of 730 “*Security Now*” podcast episodes, scraped from the public website [www.grc.com/securitynow.htm](http://www.grc.com/securitynow.htm). Data was gathered strictly for academic research purposes. Speaker demographic is predominantly homogenic (white, male, American, 30+) and consistent throughout the dataset. The texts are of *semi*-*unstructured*, *conversational* nature, and the topic range is narrow; each episode deals with issues of personal computer security, but due to the format are sometimes more private in nature. Data was scraped for the period of August 2015 to October 2019.

A. Curation Rationale • Why was the data gathered?

B. Language Variety • What language, dialect, other focus?

C. Speaker Demographic

D. Annotator Demographic E. Speech Situation • Was this a public speech? On the phone?

F. Text Characteristics • Genre, topic(s), vocabulary

G. Recording Quality

H. Other

I. Provenance Appendix

Do any of the social implications highlighted in the Hovy and Spruit (2016) bear on your project? E.g. cam you imagine any dual-use problems?

2. What are possible types of bias in your datasets, or in datasets we've worked with this semester?

3. How do you balance the potential costs/risks with the potential benefits of an NLP project? What are some best practices for minimizing bias/privacy risks/exclusion/overgeneralization?