Shomir’s Tenure Narrative Statement – 2023-2024 Cycle

1 Research

My research portfolio has a core focus on studying privacy using methods from natural language processing (NLP). Privacy is challenging for technology users, and text serves as the primary format for organizations’ data practices and as an important medium for sharing personal information. My research to identify and analyze organizations’ data practices at a large scale has created novel insights about the state of consumer privacy, as well as privacy-enhancing technologies that help users gain control over their personal data. My research on online social networks shows how users’ behaviors contrast with their sharing preferences, suggesting how best to protect and inform users. Additionally, my broader portfolio applies NLP to problems in security and social science.

In the narrative below, note that ACL, EMNLP, COLING, and EACL are top-tier publishing venues for NLP, and WebConf, TWEB, PoPETS, NDSS, UbiComp, CSCW, and CNS are top-tier for privacy and related topics.

1.1 Depth

Some strands of research in my core focus are:

* Large-scale collection and analysis of privacy texts: I led the NSF-funded team that created PrivaSeer, a privacy policy search engine and one of the largest privacy policy corpora publicly available. Our work created knowledge about the state of online privacy at a previously untenable scale and enabled the creation of human language technologies that can accurately extract information from these documents (ACL 2021, DocEng 2021/2023, ICWE 2021).

* Privacy policy question answering and choice management: We discovered latent trends in the organization of privacy policies (EMNLP 2018), showed that crowdsourcing annotations for privacy policies is practical (TWEB 2018), automated question answering for privacy policies (EMNLP 2019), created user-facing tools for opting out of data collection (WebConf 2020), surveyed the state of NLP research to benefit consumer privacy (ACL 2021), and discovered divergences between privacy policies in multiple languages (LREC 2022).

* Privacy on online social networks (OSNs): We discovered that demographic factors affect self-disclosure rates on OSNs (PoPETS 2021), with findings that inform efforts to increase privacy equity. In another project, we developed a method to automatically detect doxing, a dangerous phenomenon where an OSN user maliciously releases sensitive information about another person (CSCW 2022).

1.2 Breadth

I cultivated additional projects, including:

* Characterization of Demographic biases in large language models (LLMs): My lab’s work has shown that LLMs, which are increasingly vital for building human language technologies, exhibit harmful biases against people with disabilities (COLING 2022). Our work has also revealed a large assortment of unequal biases for and against nationality demonyms (e.g., “German”, “Malaysian”, “American”, etc.) in GPT-2, a popular LLM (EACL 2023, AIES 2023).

* Expectations versus reality for email security: I was part of a collaboration that developed a method to defend against file-injection attacks that use email searches (CNS 2019). My lab also

studied multilingual differences between collections of scam emails (SOUPS 2020 poster) and contrasts between university scam email education resources and scams received by university email addresses (SOUPS 2021 poster, WebConf 2023).
* Law enforcement language and outcomes for minorities: Our work to study transcripts of police radio communications has shown disproportionate contact between law enforcement officers and members of minorities, as well as acute privacy risks when officers describe people over the radio. We are aiming for a CSCW submission in late 2023.

1.3 Research Prior to Penn State

My research post-PhD but pre-Penn State includes:
* Foundational data-driven studies of privacy policies: I led the creation of the first large-scale annotated corpus of privacy policies (the OPP-115 Corpus, published in ACL 2016) and the first assessment of crowdworkers' accuracy in this domain (WWW 2016, Best Paper Finalist). We also developed and deployed automated methods for detecting mobile app non-compliance with privacy laws (NDSS 2017), extracted user choices from policy text (EMNLP 2017), and created an ontology for automated reasoning about data practices (SWJ 2017).
* Privacy in online social media: Users' expectations for privacy are not always met in online social media, and my work revealed this discrepancy in the contexts of post deletion (CSCW 2013) and real-time location sharing (UbiComp 2013). With collaborators I also surveyed prior work on privacy and security nudges (ACM CSUR 2017).
* Metalanguage in natural language: My work was the first to create a corpus of English metalanguage (ACL 2012), followed by automatic detection of the phenomenon (IJCNLP 2013) and a study of similar deictic phenomena in language (EMNLP 2014, GWC 2016). This strand is the only one in this narrative statement that relates to my dissertation.

1.4 Funding and Artifacts

I have been the Penn State PI or co-PI on five external awards or subawards (4 NSF, 1 NIH) totaling over $2M awarded or anticipated research funding. These include being the lead PI on the PrivaSeer NSF grant, worth $700K to Penn State and $1.2M across all sites. My funded collaborations include faculty and professionals at Carnegie Mellon University, University of Michigan, University of Chicago, Fordham University, and the Future of Privacy Forum.

My research has produced artifacts (in addition to publications) with practical value for the public, industry, and other researchers. Among them are the Opt-Out Easy web browser extension, released for Chrome and Firefox, to help internet users quickly take advantage of websites' privacy options; the PrivaSeer search engine and corpus, one of the most significant resources in this space (see description in 1.1); and the OPP-115 Corpus and Explore site, also a pair of foundational resources for the research community. Toward technology transfer, in Fall 2021, [redacted company name] and [redacted company name] bought commercial licenses to the OPP-115 Corpus. One of my postdoctoral universities holds the license, but my work as faculty led to the commercialization.

2 Teaching

2.1 Courses

My teaching portfolio at Penn State includes five courses covering the spectrum of graduate level, upper-level undergraduate, and intro-level undergraduate. One (IST 597/574) I created and established a permanent number for. Three (IST 510, IST 472, IST 110H) I was told had a troubled history before I taught them, and very few (if any) prior materials were available. All five I redesigned from scratch.

* IST 597/574, Natural Language Processing for Sentiment, Semantics, and Discourse: Each of the three topics in the course title respectively occupied a third of the semester. Each unit consisted of a lecture on background material, followed by student-led discussions of published papers. The term projects produced publication-quality results: at least three teams submitted manuscripts to peer-reviewed venues, with one acceptance and two future resubmissions.
* IST 510, Foundations in Computational Informatics: I redesigned this course with a focus on teaching students from a variety of scholarly backgrounds to think critically about computational concepts and to choose among computational tools for a variety of tasks.
* SRA 472, Integration of Privacy and Security: This was a required course for many graduating seniors in our college. Since it was primarily intended to be a privacy course, I organized it into units on privacy fundamentals, methods, and applications. I split classroom time between lectures and group discussions. I created a term project that was a two-part case study of a notable consumer privacy failure: students presented the failure and the mistakes that led to it, and then they proposed a solution to avoid the failure.
* IST 110H, Honors Introduction to Information, People, and Technology: I augmented the standard (non-Honors) course objectives with material on critical thinking and technology ethics. Classroom time focused on discussions of assigned readings, with some lectures and group activities mixed in. The depth of the classroom discussions and the creativity of the term projects suggested positive outcomes.
* PSU 17, First-Year Seminar College of Information Sciences and Technology (for data science majors specifically): I taught this course on overload three times. I modeled it after IST 110H while devoting a greater fraction of time to resources to succeed in college.

In 2021 I led the development of a proposal for AI 430 (Text and Natural Language Processing), as part of the Data Science area faculty's effort to develop a curriculum for a new AI major. In January 2022 the proposal was approved by the area faculty.

I adjusted to remote teaching during the pandemic by frequently using Zoom breakout rooms for class discussions: I gave students a question to focus on, sent them to breakout rooms for a limited duration, and afterward led a larger discussion of their answers. In Spring 2021 I asked an Instructional Consultant from the Schreyer Institute for Teaching Excellence to observe my class. Her feedback was supportive of my discussion strategies, how I structured class time, and my classroom demeanor (“enthusiastic” and “welcoming”, from her enclosed letter). She recommended that I perform a mid-semester evaluation and debrief students about the results, and I implemented that advice.

2.2 Student Mentorship

As of Summer 2023, I supervise four PhD students. One has successfully proposed his thesis topic and all have passed their qualifying exams. Three have one or more publications in high-impact venues.

I have graduated three MS students at Penn State, and two have begun jobs as a Software Engineer at Apple and an Associate Application Developer at ADP. I have also supervised 14 undergraduate researchers, and highlights of their work include an Erickson Discovery Grant, first author on conference publications (WebConf 2023, JURIX 2020), coauthor on a journal paper (PoPETS 2021), and coauthors on two poster abstracts (SOUPS 2020/2021).

Just over half of my research mentees at Penn State have been women, showing my support for gender diversity in informatics. Additionally, to support inclusion of many underrepresented groups, I wrote Advice for Students (https://shomir.net/advice_for_students.html), a set of guides for typically unwritten norms in academia. They received enthusiastic feedback from faculty at Penn State (at least two IST faculty share these guides with their classes), from faculty at other universities, and from students.

To improve my skills for working with students, I attended seminars on mentoring graduate students and helping students in distress. I also attended Penn State Student Affairs' two-part LGBTQ+ inclusiveness seminar and a QPR training seminar for suicide intervention.

3 Service

3.1 College and University Service

In Fall 2020 I arranged for the International Association of Privacy Professionals to annually give a Westin Award ($1K cash prize plus other benefits) to a Penn State IST student who is interested in consumer privacy. I organized the selection process and continue to be involved in it. My service to the college includes three faculty hiring committees, three qualifying exam committees (once as chair), a term on the graduate recruitment committee, and a term on the awards committee. I also served as an interviewer for the university's Millennium Scholars program every spring since starting at Penn State, and I co-organized the IST-EECS NLP Colloquium Series every fall since 2019.

I also represented our college externally. I serve on the University Faculty Senate, and I represented IST at NSF workshops to create Departmental Plans for Broadening Participation in Computing in 2019 and 2022.

3.2 External Service

In 2019 I organized the AAAI Spring Symposium "Privacy-Enhancing Artificial Intelligence and Language Technologies" (PAL), which featured two keynote speakers, presentations of peer-reviewed submissions, and discussion panels. This was the successor to a AAAI Fall Symposium that I also led in November 2016. I have also held leadership roles at other events, including leading breakout sessions on consumer privacy at NSF SaTC PI meetings in 2019 and 2022 and co-leading a group mentoring session at ACL 2020. Toward peer review, I continue to serve on the editorial board for the Journal of Intelligent Information Systems and as a standing reviewer for Computational Linguistics. I also served on the program committees for top NLP conferences, including (most recently) ACL 2023, EMNLP 2022, and AAAI 2022, and I reviewed grant proposals for the National Science Foundation and for other funding agencies.

Companion Notes for Shomir's Tenure Narrative Statement – 2023-2024 Cycle

Overview

I am sharing a lightly edited version of my tenure narrative statement for the benefit of junior faculty. I provide these companion notes and the statement as-is with no guarantees for their suitability for any purpose. In addition to reading these materials, I recommend speaking with other faculty who can provide further advice for your individual circumstances.

Context

I am a computer scientist by training, which is important for interpreting my narrative statement. Within computing, conferences are often the top-tier publication venues and faculty are expected to get external funding to support PhD students. My advising style is to be a research director rather than a hands-on researcher, and because of that my publication rate began low (when I had few advisees) and increased over time.

Twelve years passed between when I received my PhD and when I filed for tenure. Prior to becoming faculty at Penn State, I held multiple postdoctoral positions and spent two years as an assistant professor at a different university, though I reset my tenure clock to zero when I moved. By the time I arrived at Penn State I already had substantial experience with grant writing, teaching, and advising student researchers. Assistant professors who begin immediately after their PhDs need more time to develop these skills, and my impression is that promotion and tenure committees will (should) adjust their expectations.

Structure

I organized my narrative statement around the standard triad of responsibilities for tenure-line faculty, with sections for research, teaching, and service. When I started at Penn State, the dean of the college (who has since left) said that pretenure faculty were expected to spend 40% of their time on research, 40% on teaching, and 20% on service. I approximately mirrored that distribution in my narrative statement: 47% of the word count is research, 38% teaching, and 15% service.

I used the same structure for my second-year and fourth-year review narrative statements. To create each subsequent narrative statement, I heavily edited the previous one. My impression is it's not obligatory to start from scratch, but you can if you want to.

Miscellany

- Early in the narrative statement I pointed out the value of conferences as publishing venues in my research areas, in case reviewers were unfamiliar with that practice (i.e., if they were more familiar with journals as top publishing venues).
- Lists felt inevitable when I was writing, but I tried to give them meaningful content. Instead of listing publications, for example, I listed research strands and then described each one.
- I've heard faculty say that it's necessary to graduate a PhD student to get tenure, but I didn't do that. Instead I documented my current PhD students' progress toward completing the degree requirements (§2.2).

- I mention in §1.4 that two companies purchased licenses to a dataset that I created as a postdoc at Carnegie Mellon University. CMU's technology transfer office didn't have permission to make those purchases public, but they OK'd the company names appearing in my narrative statement, on the basis that the readership wouldn't be the public. I've redacted their names in this copy.
- I explicitly identify the research that I did after my PhD but prior to Penn State (§1.3) to avoid any misunderstandings about when I did things, while still including them so I would get some credit for them. Also, to emphasize my growth as a researcher, I pointed out that only one of my post-PhD research strands was connected to my dissertation.

Further Reading

For more advice on being pretenure faculty, see Tenure Worth Wanting and the other "Academia as a Career" guides on my advice page.