Max Klein: Statement of Purpose

University of Minnesota Computer Science Department

Research agenda

Gender, race, sexuality, nationality, social-class, native-language, height and weight. Can we make an exhaustive list of social biases? I believe we already have, but haven't realized it yet.

Accounting for social biases is a main barrier to entry and expansion in sociotechnical systems[1]. I'm tantalized by the idea that we could automatically detect biases we haven't yet identified – that we could uncover Rumsfeldian unknown unknowns. I propose a research agenda to classify the already-known social biases by as they appear in the collaborative technologies, and then search for unidentified biases using those classifications. As an explanatory example, create a statistical model of how the known skewed distributions of gender, race, and nationality exist in Wikidata (the free knowledge base that feeds Wikipedia), and then inspect *all* the property distributions to match the biased patterns. The project grows more complex by allowing property-pairs (e.g. gender by race), different sociotechnical communities (e.g. Freebase, Git-hub), and different models of bias (e.g. editorship-measures). If successful we will find overlooked stigmas of people using technology that haven't been identified yet in any other way.

Preparations

This research agenda is the direct consequence of the experience I have gathered in my academic, work, and personal portfolio. They are the three strands which constitute the braid of my readiness: technical head-down-ness, Open Culture social awareness, and self-driven curiosity.

In Spring of 2015, I will present my research on "The Virtuous Circle of Wikipedia"[2] at Computer-Supported Collaborative Work conference 2015. The paper provides a new definition and measure for the "collaborativeness" of a socio-technical community – the degree to which being a good user is correlated with editing alongside other good users. In this case I studied the non-profit Wikipedia, using economic insight and a variant of the Google PageRank algorithm. We found that the editors in the category "US Military History" was the most collaborative, but the editors of category "Sexual Acts" only seemed to edit-war. This will guide my proposed research agenda because it will help answer the criticism of whether found biases reflect society at large, or only the biases of the editing community. It will answer that question by determining whether found biases are correlated to collaborativeness.

Moreover my "Virtuous Circle" research has prepared me by stretching my technical strengths and weaknesses. The project was a collaboration with UC Berkeley School of Information PhD Thomas Maillart, who pushed me to grasp new mathematical models and higher methodological rigor. My BA in mathematics was useful in learning the required network science quickly, shown in that an early stage of the research was presented as a poster to the NetSci 2014 conference[3]. Likewise I was also pushed to come to speed with a base computational social science literature. An example of exceeding this goal is evident in the "Method of Reflections" technique which we borrowed and utilized in the paper – after which I created the first open-source implementation of the method[4].

To open source that algorithm was instinctual because I am a part of the academic-hacker Wiki Research community. It was there that my blog posts on the Gender Biases in different Wikipedia languages was picked up by Hanyang University Sociology Professor Piotr Konieczny. We came to work together on creating "Wikipedia Gender Inequality Index" (WIGI[5]), an upcoming Open Dataset of extracted information from Biography articles across all Wikipedia Languages. My contribution was to put together the technical

1 Halfaker, Geiger, Morgan, and Reidl. The Rise and Decline of an Open Collaboration System (2013)

2 Klein and Maillart The Virtuous Circle of Wikipedia (2014)

3 Klein and Maillart Poster - The Virtuous Circle of Wikipedia (2014)

4 Klein Method of Reflections Explained and Exampled (2014)

5 Klein and Konieczny WIGI: Wikipedia Gender Inequality Index

infrastructure to re-index, analyze and display the huge dataset each month. I also came to sharpen my statistical testing from Piotr as we first analyzed the predictive power of the data by Date and Place of Birth, Ethnicity, Citizenship, and Language. Currently in the submission process, WIGI is an existing prototype and first step in the stream of the research agenda put forth.

My preparedness is anchored by a final personal factor: the philosophy to steer my own course, navigating by natural curiosity. One telling fact about the above two projects is that I have not done them under the direction of an institution, either academic or industry. Both have come voluntarily without pay and on my free time. The last major directives I received were 3 years ago at the beginning of my stint as Research Assistant at OCLC Inc. where I was hired to improve Wikipedia-Library integration, which I accomplished by way of writing Wikipedia bots to add content from Library databases. After publishing about the process[6] and finally amassing over 2 million edits, I came to see what was driving me in the position. It was the Open Notebook Science - the source, data and analysis syncing publicly online as it evolves in front of me, thereby submitting my ideas to the full scrutiny of the internet hivemind. More and more my ideas spilled over the scope of the OCLC Research blog[7], and I started hosting them myself at *notconfusing.com*[8]. My natural inclinations to investigate on my own started receiving attention and energizing feedback – particularly on posts exploring the data of gender[9] and language[10] [11]. Although I didn't fully realize it then, this new networked, boss-free but still peer-reviewed world meant that I already started to pursue my PhD.

Faculty Interest

No blog is an island. As my research efforts intensify I have a growing need for guidance in methodological framing and focus. In a conversation with Brent Hecht – the faculty member with whom I'm interested in working with, I was aroused by the discussion of his research with WikiBrain around online effects geography and language (my review[12]) and overall the dedication to making social human-computer interaction statistics less difficult to compute. This ease in computation is particularly important when we consider the intensive methods employed by HaiYi Zhu. Her investigations into “shared leadership”[13] intrigue me from the machine learning perspective and also the interdisciplinary bridging of leadership theory. They harp on exactly the sort of social collaborativeness questions that I'm dancing around in my research – but that I need help in refining. Along with a veteran of that team, Loren Terveen, I would ask that they could help train me in defining and attacking the more poignant sociological questions that I do not quite know how to ask yet.

Earning my PhD from UMN, in the grand scheme, is a waypoint on the path towards becoming a professional researcher. The end product of my education is necessarily full time research, because the end goal of my education is knowing how we are effecting sociotechnical systems with outmoded social biases, which will forever be less than fully understood.

---

6 Klein and Kyrios VIAFbot and the Integration of Library Data on Wikipedia (2013)

7 Klein HangingTogether.org Blog Posts (2012-14)

8 Klein Notconfusing.com Research Blog Posts (2012-14)

9 Klein Sex Ratios in Wikidata Part III (2014)

10 Klein Actionable Metrics for Uganda and Côte D'Ivoire (2014)

11 Klein The Most Unique Wikipedias According to Wikidata (2013)

12 Sen, Li and Hecht WikiBrain Democratizing Computation on Wikipedia (2014)

13 Zhu, Kraut, Kittur Effectiveness of Shared Leadership in Online Communities (2012)