Social Data Science: Community-Centered Quantitative Methods for Trans Populations

According to a widely-circulated 2015 statistic, *the life expectancy for an African-American transgender (trans) woman in the U.S. was only 35 years*. Compared to the overall life expectancy at birth in the United States in 2015 (78.8 years), this statistic indicates that African-American trans women would be expected to live less than *half as long*. It was also one of the few pieces of quantitative data cited in reports on the murders of trans women of color in the United States at this time. At face value, this statistic is striking because it paints a disturbingly bleak picture of the de facto right to life of African-American trans women. Considered with its story of provenance and use, it becomes striking for an additional reason, as it illustrates the near nonexistence of scientific knowledge about trans populations.

A small but growing body of knowledge speaks to the *material* adversity—e.g. economic, social, and medical disenfranchisement—faced by trans women of color and the broader trans population in the U.S., although little research, if any, represents population-level statistics that are comparable across populations. The life expectancy statistic discussed above is a notable exception. However, tracking the statistic to its source does not inspire confidence. Instead, it highlights an additional kind of *informational* adversity, where trans populations lack the scientific data and methods to make sense of their experiences in broader social contexts and advocate for themselves in our increasingly data-driven world of policy and practice. The statistic originated in a report on violent crime in the Americas by the Inter-American Commission on Human Rights[1] and did *not* represent life expectancy. Instead, it *became* a life expectancy statistic through a process of misunderstanding and distortion as it moved across pop culture media outlets, social media, and blog posts, reminiscent of a game of "telephone"[2]. Why did a fallacious "folk statistic" become the quantitative cornerstone of public discourse around trans lives and advocacy in 2015? The spread of this unfounded life expectancy statistic demonstrates the importance of quantitative work in rendering marginalized experiences visible to the broader public. The ability to locate oneself in the world of quantitative data affords access to a shared language that is often the basis of policy, activism, health interventions, and making sense of one's own experience. This life expectancy statistic is incorrect but it's all we—social scientists and trans people—have. What this statistic best represents is a breakdown in our public knowledge systems for trans populations; it is a statement of need. Yet to reject it is a dismissal of what, by all accounts, is an epidemic of violence.

Intellectual Merit. It's informative to compare the trajectory of this statistic to the research infrastructures that support "mainstream" social science knowledge. Across many fields, including sociology and demography, data science methods are evolving in new, exciting, and often "invisible" ways as they draw on developments in data science. Rather than being visibly constituted by online crowd work, the processes that produce life expectancy statistics for most

[1] See http://www.oas.org/en/iachr/media_center/PReleases/2014/153.asp and https://www.oas.org/en/iachr/lgtbi/docs/Annex-Registry-Violence-LGBTI.pdf

[2] "Telephone" is a game where a phrase is passed person-to-person through a group of people.

cisgender Americans are largely invisible, reflecting a relatively functional infrastructure for scientific knowledge. As interlocking digital systems increasingly become the basis for computational social science—such as the data pipelines that span census form responses and the statistics on the CDC website—it is a moral necessity that the data flowing across these systems equitably serve the informational needs of marginalized groups. This work will further empirical sociological work by developing data science methods that more accurately and ethically represent social identities.

Broader Implications. Not only is further empirical work on trans populations needed, but there are severe limitations of current quantitative methods for representing the experiences of trans people with respect to gender category. These limitations lie within a broader set of theoretical and methodological questions surrounding how scientists can produce knowledge as a public resource for vulnerable populations. To understand and communicate their experiences, marginalized groups need to be visible—more than footnotes—in scientific models of the past, the present, and the future. In the broader context of the informational and sensemaking needs of underrepresented groups, this work will construct new pathways for social scientists to empower rather than exclude marginalized groups in scientific knowledge production.

Focusing on trans identities, the proposed work aims to bridge the gap between marginalized groups and sociological work by developing community-centered data science methods, where "community-centered" refers to considering target communities to be experts of their own experiences and needs. There are three proposed steps:

1. Contextual Empirical Work. This step will involve empirical examination of social media trace data (e.g. from Reddit, Tumblr, and personal blogs) and archival/narrative data (e.g. from the Digital Transgender Archive) to surface themes of how trans people have historically self-selected identity categories. Qualitative text analysis will be used to identify patterns of self-categorization and network analysis will be used to identify contextual and relational attributes of identity categories. The output will be a broad relational/genealogical understanding of gender categories.

2. Participatory Design of Methods. The next step will be to use co-design sessions to examine how gender identities can be used in quantitative research. Participants will be presented with a question about a gendered experience (e.g. how do you experience your gender on public transit) and will co-design a quantitative representation of their experience. An additional second stage of qualitative analysis will be conducted on transcripts and artifacts created during the design session. The output of this step will be a rich, focused analysis of how individuals classify and quantify their gender in a given context.

3. Extension, Evaluation, and Critique. This final step will synthesize the results of (1) and (2) into a data science methodology and apply it to a new context. The goal of this final step will be to assess the successes, limitations, and ethical implications of the methodology. The output will be a quantitative, community-centered methodology and a critical evaluation of the methodology compiled while applying it to a research question.