Serina Chang — Research Statement

My research develops computational methods to tackle complex societal challenges, from pandemics to polarization to supply chains. I leverage **novel data sources**, such as mobility data and search logs, to capture **human networks** and **behaviors** at the center of societal systems. To address challenges that arise from real-world data, I develop new methods blending **machine learning (ML)**, **network science**, and **natural language processing**. I use these methods to derive policy insights and build decision-support tools, paving a new way for high-stakes decision-making powered by large-scale computation and data.

In an interconnected and fast-moving world, effective policymaking increasingly relies on understanding complex human networks and behaviors. For example, pandemic response requires understanding how disease spreads through contact between individuals and how individuals alter their behavior in response to policies and disease. Novel, high-frequency data sources, such as cell phones and search engines, introduce new opportunities to capture these networks and behaviors at scale. However, there remain **substantial gaps** between real-world systems and what is recorded in data, which preclude our ability to fully leverage such data for critical decision-making. My research seeks to close these gaps by (1) developing methods to **infer robust signals about real-world networks and behaviors from novel data sources**, and (2) leveraging these signals to build **decision-support tools for policymakers**.

Specifically, my work thus far has addressed three fundamental data challenges. First, physical human networks are *rarely known*, but they are essential for downstream policy problems. I have identified novel data sources, such as mobility data, that provide aggregated views of these networks, and developed statistical methods to infer fine-grained networks with local spatiotemporal patterns that greatly impact health outcomes and disparities [6-8,13][1]. Second, novel data provides a wealth of new information but is often *unlabeled* for signals of interest (e.g., user intents). I have built ML systems to efficiently label and organize such data (e.g., into ontologies) while balancing user privacy, such as through anonymized search logs [12], social media [2], news articles [3], and public speeches [10]. Finally, even when networks are observed, the *underlying mechanisms* that drive their evolution remain unknown. I have developed methods blending causal inference and graph ML to discover mechanisms driving dynamic networks, such as spillover effects in mobility networks [11] and production functions governing supply chains.

My research has made **important contributions in computer science** (CS), as recognized by my publications in top CS venues (KDD, AAAI, ICWSM, EMNLP, EACL), PhD fellowships (NSF GRFP, Meta), and awards (EECS Rising Stars, Rising Stars in Data Science, Cornell Future Faculty Symposium). My work also has **broad impact on policymaking**, particularly in public health. My first-author paper in ***Nature*** [6] received coverage from over 650 news outlets, and governments around the world used our results to shape their pandemic policies. I also developed a decision-support tool for the Virginia Department of Health [7]; this paper received the **Best Paper Award in KDD** (Applied Data Science Track) as the single best paper out of 705 submissions. Finally, my work is **highly interdisciplinary**, with publications in top scientific journals (*Nature*, *PNAS*, *Physical Review Research*) and collaborations with experts across fields, including public health, epidemiology, sociology, economics, and social work.

Inferring fine-grained mobility networks for pandemic response

Infectious diseases spread through human contact, but real contact networks are typically unknown. Prior work relies on imperfect alternatives, such as generating synthetic data or using coarse-grained networks (e.g., airline networks). During the COVID-19 pandemic, we proposed a new solution: **to infer fine-grained mobility networks from aggregated cell phone location data**, by formulating a general network inference problem and developing statistical methods to solve it.

***Network inference*.** The objective of our *network inference problem* is to infer a dynamic network from its 3-dimensional marginals, i.e., its time-varying rows, time-varying columns, and time-aggregated interaction matrix. This problem appears across domains such as mobility, transportation, and migration,

[1] Citation numbers reference to the numbered publication list in my CV.

where it is more feasible—due to privacy or real-time constraints—to collect the marginals of a network than its time-varying interiors. To solve this problem, we repurposed the *iterative proportional fitting* (IPF) procedure, which has been studied extensively in the context of matrix balancing, but not for network inference. In our work, **we established the first theoretical basis for using IPF to infer dynamic networks from their marginals** [13]. Specifically, we motivate the use of IPF by designing a generative network model whose parameters are properly estimated by IPF, since its maximum likelihood objective is dual to the Kullback-Leibler divergence minimization problem implied by IPF. Our model provides a principled framework for using IPF in this setting, revealing implicit assumptions and connecting it to Poisson regression, yielding results about IPF's robustness to data noise and sparsity. By applying IPF to aggregated location data, we inferred hourly mobility networks from census block groups (CBGs) to individual points-of-interest (POIs) (Fig. 1). IPF is computationally lightweight, allowing us to infer networks covering **98 million people with 5.4 billion hourly edges**. We also released our networks, which have been downloaded by hundreds of researchers and used in other studies.

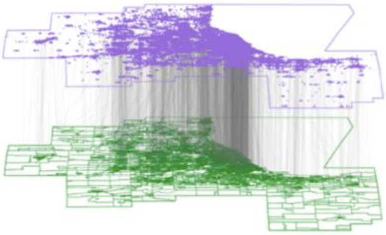

Figure 1. Inferred mobility networks between CBGs (bottom) and POIs (top).

***Insights on reopening and disparities*.** By designing a new epidemiological model that integrated our mobility networks, we could realistically simulate **who was infected where and when** down to the POI and hour. Even though our model only had three free parameters, our IPF-inferred networks enabled the model to accurately fit daily COVID-19 case data in the US. The granularity of our networks also allowed us to analyze **heterogeneities in risk across POIs and CBGs**. Based on our model, we found that reopening with targeted restrictions could limit new infections without shutting down the entire economy. Our model also correctly predicted higher infection rates among lower-income and less white CBGs, solely from mobility patterns. Two mechanisms explained these predicted disparities: disadvantaged groups were not able to reduce their mobility as sharply, and even within the same category, the POIs that disadvantaged groups visited were more crowded with longer visits, thus associated with higher risk.

Our paper in ***Nature*** [6] received widespread coverage from over 650 news outlets, including an interactive article in *The New York Times* (NYT), with currently the **10th most online impact** among 94,600 papers ever published by *Nature*. I was invited to present our findings to several US state and county governments, at an OECD seminar, and in guest lectures at Stanford, Cornell, and Columbia; our work is also taught at other schools (e.g., NYU, UNC Chapel Hill). Our work has over 1,200 citations to date, and was also cited by the CDC, *NYT* Editorial Board, US Joint Economic Committee, in two briefings for the US Supreme Court, and by governments (e.g., US, Canada, Poland, Japan) when announcing public health policies.

***Tools for policymakers*.** During the pandemic, policymakers had to make difficult decisions about mobility restrictions, which were a primary intervention for controlling transmission but also placed heavy burdens on businesses and individuals. Our model, with its ability to quantify both predicted infections and visits, presented a unique opportunity to support policymakers' needs. In collaboration with the UVA Biocomplexity Institute, I built a **decision-support tool** based on our model for the Virginia Department of Health (VDH). Our tool allowed them to assess tradeoffs between mobility and predicted infections through thousands of possible reopening plans. Building this tool required many extensions to our model, including a computational infrastructure to deploy the model at scale, compressing 2 years of compute time into a few days, and a new model dashboard that we designed with VDH's feedback (Fig. 2). The advancements in this work allowed us to transform our model into a deployed tool for public policy, for which we received the **KDD 2021 Best Paper Award** in the Applied Data Science Track [7]. We also aided VDH on vaccine distribution, delivering **weekly recommendations for vaccine site placements** that prioritized undervaccinated populations (e.g., Black and Latinx) based on aggregated mobility patterns [8].

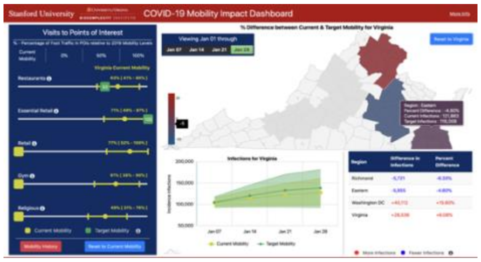

Figure 2. Our dashboard helps policymakers to quantify tradeoffs in reopening policies.

Extracting real-world signals from unlabeled data

Novel data provide vast amounts of new information, but they are often unlabeled. To address this challenge, I have developed ML systems to efficiently label and organize such data into **precise real-world signals** that can guide policymaking and yield valuable societal insights.

Vaccine hesitancy & search logs. In collaboration with Microsoft Research, I leveraged billions of anonymized search logs to derive precise signals about vaccine trends [12]. First, we developed a semi-supervised, graph neural network-based classifier that accurately detects when a user is seeking the COVID-19 vaccine on Bing. Our classifier uses a novel pretraining objective, based on personalized PageRank, to achieve strong performance in all US states (AUCs > 0.9). **Our vaccine rate estimates agree strongly with CDC vaccine rates** ($r > 0.86$) with our search signals preceding the CDC by 7-15 days.

With our classifier, we could estimate vaccine rates to the level of ZIP codes, 10x the granularity of CDC data. We released these estimates publicly, as the **most comprehensive dataset of ZIP-level COVID-19 vaccine rate estimates** to date. Our estimates reveal substantial heterogeneity in rates within counties (Fig. 3), motivating the need for more fine-grained data to support policy (e.g., for vaccine site placement). We also used our classifier to identify vaccine early adopters and holdouts. We found that holdouts, compared to matched early adopters, were 69% more likely to click on untrusted news. Furthermore, we discovered vaccine concerns directly from their clicks, by organizing 25,000 vaccine-related URLs into a **novel ontology of vaccine concerns**, which we also released. We found significant differences in the concerns of holdouts versus early adopters, as well as varying concerns *within holdouts* across demographics and over time, revealing the need for tailored interventions that address individuals' specific concerns, instead of one-size-fits-all solutions.

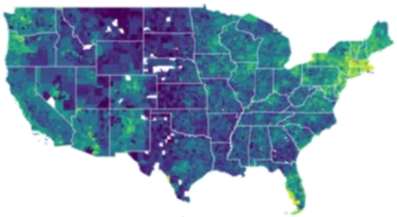

Figure 3. Estimated vaccine rates per ZIP.

Immigration & political speeches. In work published in *PNAS* [10], we used **large language models** to characterize sentiment towards immigrants from 140 years of US political speeches. We trained the models to identify precise segments of speeches that were about immigration, and to classify their stance as positive, neutral, or negative. Furthermore, we developed unsupervised methods to uncover *linguistic framings* of immigrants, developing lexicons for 14 frames (e.g., family, crime), and to measure *implicit dehumanizing metaphors*, with a novel language modeling technique based on masking mentions of immigrants.

Our methods enabled the **largest quantitative analysis of political attitudes towards US immigrants**. We found that average attitudes are more positive now than ever, but political parties have diverged to the extent that Republican speeches today are as negative as the average speech in the 1920s, an era of strict immigration quotas. Furthermore, modern Republicans are significantly more likely to frame immigrants in terms of crime and legality and to use language suggestive of dehumanizing metaphors. We also observed a striking similarity between the framings of Mexican immigrants today and Chinese immigrants during the period of Chinese exclusion. This work, which has been taught at schools including Stanford, Cornell, USC, and University of Buffalo, reveals lasting antiimmigrant sentiment that echoes past exclusionary policies, despite the elimination of race-based restrictions and more positive attitudes overall today.

Uncovering causal mechanisms of dynamic networks

Even when networks are observed, the underlying mechanisms that drive their evolution remain unknown. To learn these mechanisms from network data, I develop methods blending causal inference and graph ML.

Estimating policy spillovers. In our AAAI'23 paper [11], we developed a causal framework to estimate *geographic spillover effects* from mobility networks. Such spillovers may occur when nearby regions have differing policies and populations travel to circumvent local restrictions, undermining their intended effects. However, estimating causal effects of policies from observational data is challenging, due to potential confounders. In our work, we identified a pandemic policy in California that determined each county's restrictions based on thresholds of COVID-19 metrics, lending to a regression discontinuity design-based framework to estimate **unconfounded spillover effects**. With this framework, we quantified the cost of spillovers on local policies, finding that county-level policies were only **54% as effective** as statewide policies at reducing mobility. However, we showed that strategic macro-county restrictions—where county groupings were optimized by solving a graph partitioning problem—could recover over 90% of statewide reductions, balancing competing objectives of policy efficacy and flexibility between jurisdictions.

***Production learning on supply chains*.** Supply chain disruptions have enormous societal costs, partially due to how shocks can propagate over supply chain networks, similar to how diseases propagate over contact networks. However, a unique characteristic of supply chain propagation is that when a firm experiences a shortage in input products, only its output products that rely on those specific inputs are affected; thus, in order to accurately model propagation, we need to learn these underlying *production functions*. In an ongoing collaboration with Hitachi, Ltd., we have developed the **first ML model for this new setting**, by combining temporal graph neural networks with a novel inventory module and product-product attention weights, inspired by neural models for *temporal causal discovery*. Our model jointly infers each firm's internal production function and predicts future transactions, outperforming strong memorization-based baselines on a supply chains dataset with billions of transactions from 2019 to present.

Vision for future research

My future work will continue at the **intersection of novel data, networks, and policy**, evolving with the advent of new data, computational capabilities, and policy needs. Directions I am excited about include:

***Graph ML for complex human networks*.** Graph ML has seen a recent explosion of interest in temporal/dynamic graphs (e.g., see the Temporal Graph Benchmark from 2023). Complex human networks are temporal and introduce further challenges, such as privacy constraints, missing or biased data, and strategic behavior such as spillovers. Furthermore, human networks are often missing from graph ML datasets, due to challenges in collecting large-scale network data with rich node-level features per individual. I am interested in exploring privacy-preserving techniques to construct large-scale human networks that can be released as benchmarks; to this end, I am already collaborating with Hitachi to generate and release realistic synthetic networks based on their proprietary supply chains data. I am also interested in developing new graph ML models that address the unique challenges of complex human networks. These models would both push forward the frontier of graph ML and power advances across policy domains, such as public health or online social platforms, where challenges are driven by complex human networks.

***Large language models for social simulation*.** Large language models (LLMs) exhibit unprecedented capabilities; in particular, recent works have shown their potential for simulating human agents [Park et al., 2023]. In ongoing work, we have also shown promising results for using LLMs to generate social networks, showing that they replicate known results on network homophily. I hope to greatly expand on this work to develop large-scale social simulations with LLMs that replicate complex societal phenomena, such as polarization, misinformation, and pandemic behaviors. Creating robust simulations will require new methods to address the limitations of LLMs; for example, their difficulties with graphical reasoning, risks of demographic biases, and sensitivity to wording in prompts. These methods would enable more powerful and flexible social simulations, which could be used broadly by social scientists and policymakers.

Direct support for policymakers. I will also continue to directly support policymakers through decision-support tools and data analysis. The COVID-19 pandemic revealed the usefulness of novel data for understanding population health and behaviors, and these lessons generalize to other crises, such as climate-based risks. For example, mobility data could help to identify populations at-risk for extreme heat, and search logs could reveal population needs after natural disasters. I hope to contribute to these efforts, building on my past work leveraging novel data and my relationships with public health departments.

Given its interdisciplinary nature, my research is highly fundable by public, philanthropic, and industry sources, as evidenced by my fellowships from NSF, Meta, and Stanford, and a grant proposal I wrote, awarded by NSF for $850,000, on computational methods to incorporate behavior into epidemiological models. The directions above capture the promise of my research to continue making high-impact contributions, by developing cutting-edge computational methods to meet pressing policy needs.