Preliminary Modeling for Anticipating Impacts with Stakeholders for Participatory AI Projects

MICHIGAN ENGINEERING
COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF MICHIGAN

Serafina Kamp, Danniell Hu, Jowaria Khan, Ben Fish, Liz Bondi-Kelly

University of Michigan – Ann Arbor, Department of Computer Science Engineering

realaize

Motivation

Idea Generation → (Anticipate Impacts; AI System Description) → *Deployment*

Key Question: *How should we balance **resource constraints** and **discovering unknown impacts** in this loop?*

Example: *PFAS Prediction in Water Sources*

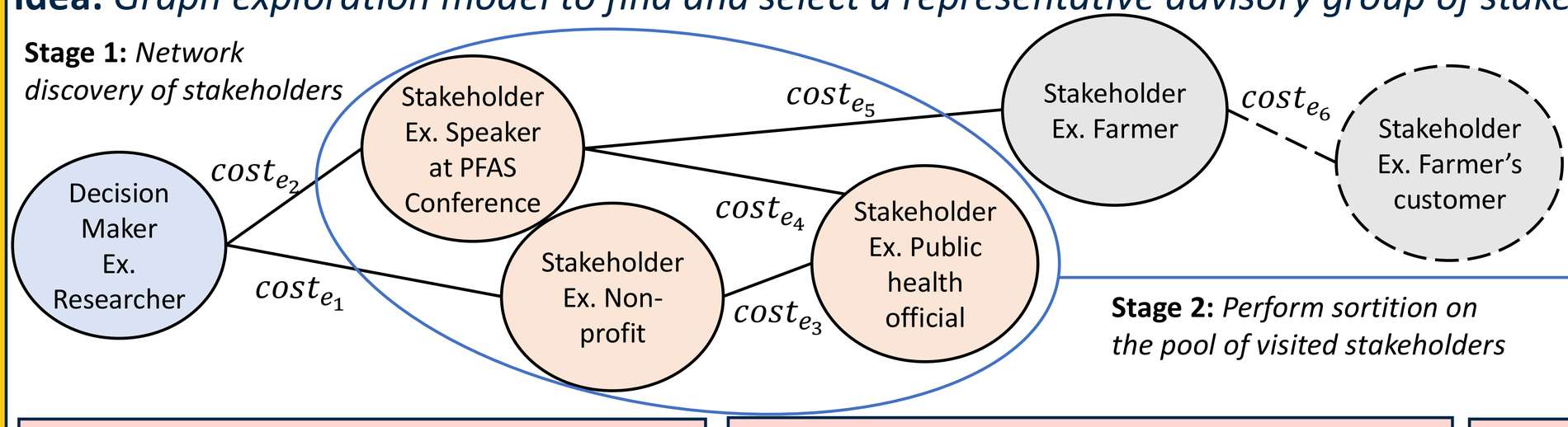
Cold emailing to connect with environmental non-profit. Initial idea generation with them. → (National PFAS conference (Economic impact to farmers), Water Center conference (Health impacts in communities), Advisory group meetings (Model prediction visualization impacts), *Additional anticipation methods? (other impacts?)*; FOCUS: Hydrology-Informed, Noise-Aware PFAS Contamination Mapping From Satellite Imagery (Khan et al., 2026)) → *Deployment TBD*

Khan, Jowaria, et al. "FOCUS: A Noise-Aware Geospatial Learning Framework for PFAS Contamination Mapping." *4th ICLR Workshop on Machine Learning for Remote Sensing (Best Paper), 2026*.

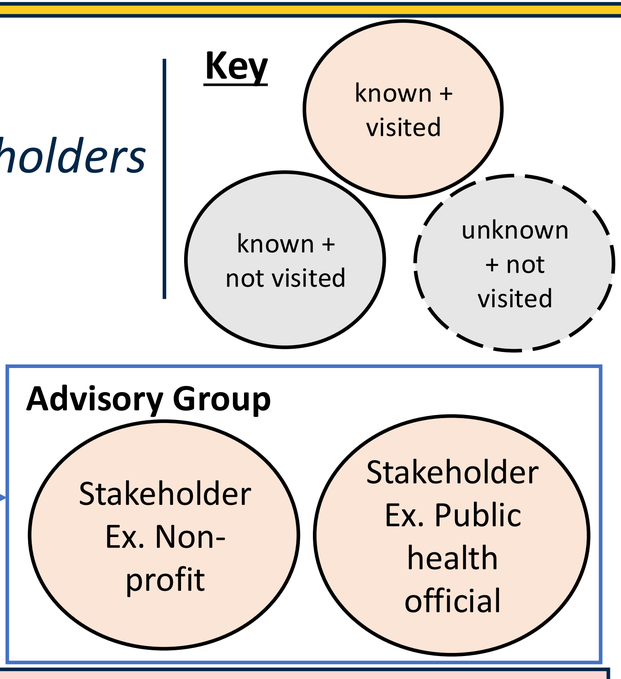
Modeling Attempt and Pitfalls

Idea: *Graph exploration model to find and select a representative advisory group of stakeholders*

Stage 1: *Network discovery of stakeholders*



Stage 2: *Perform sortition on the pool of visited stakeholders*



Pitfall: *Estimating beliefs about impacts based on network position and defining an access mechanism to observe them **requires significant assumptions***

Pitfall: *Approaching individual stakeholders to ask for their impact beliefs in exchange for a chance to be in an advisory group **imposes unnecessary burdens***

Pitfall: *Defining and selecting a representative group **requires initial knowledge of what features can be found in the network***

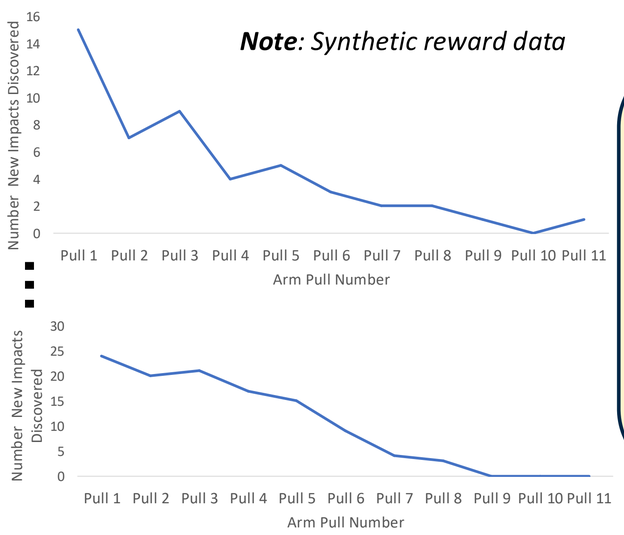
Proposed Model and Open Questions

Idea: ***Bandit model** for **choosing stakeholder engagement methods** to **discover potential impacts** with stakeholders connected to your established network of community partners*

Learning Model: *Stochastic non-stationary bandit with a knapsack*

Decision Maker Ex. Researcher + community partner team

- Engagement method 1, $cost_{em_1}$ Ex. Run a survey with partner
- **Arm pull:** Stakeholders connected to network voluntarily engage with method Ex. Responding to an email sent out to a listserv, attending a workshop held in town
- Engagement method k, $cost_{em_k}$ Ex. Host a design workshop



***Note**: Synthetic reward data*

Rewards: Each pull of an engagement method draws a sample from Θ_{em}. Given assumptions, number of new impacts discovered *decreases in expectation* with additional samples.

Assumptions:

- Researcher has a current AI system description, an established network of partner organizations, and an established set of engagement methods
- Number (and content) of potential impacts is *initially unknown* and there are *finitely many*
- For each engagement method, em, there is a (latent) fixed probability of independently discovering each impact. The (latent) joint distribution over discovering any set of impacts is denoted Θ_{em}.

Data Requirements:

- Assume natural language feedback from each engagement method can be coded (e.g., researchers and community partners come to a consensus) to categorize potential impacts
- Cost of each engagement method is known
- Fixed budget constraining number of arm pulls is known

Open Questions:

- *Reward functions related to saturation*
  - Framework is analogous to *new species discovery in ecology*, can we adapt the statistical estimates from this literature in a *decision-focused learning approach* to estimate the reward functions while learning which arm to pull?
- *Limited arm pulls and potentially delayed feedback*
  - Can we *work with community partners* to define correlations between arms or otherwise provide the model with prior information or additional signals about rewards?