Reaching a Consensus in Predictive Loops

Jiduan Wu,[1,2] Rediet Abebe,[1,3,*] and Celestine Mendler-Dünner[1,3,*]

1. Max Planck Institute for Intelligent Systems, Tübingen, and Tübingen AI Center
2. ETH Zurich
3. ELLIS Institute Tübingen

* joint supervision

Arxiv preprint presented at NetSci: `https://arxiv.org/abs/2603.12137`

In many real-world systems, machine learning models do more than describe outcomes. They shape user experience, decisions, and incentives, and thereby help create the outcomes they aim to describe. When these models are then repeatedly updated based on fresh observations, this induces a feedback dynamic where predictions and outcomes shape one another. We refer to this as the predictive loop. In this work we study how such predictive loops alter the flow of information in social networks.

We work with a simple assumption that a platform's predictions influence individual opinions, which then evolve through peer interactions and form the training data for future platform model updates. We demonstrate that this co-evolution can induce a novel equilibrium that qualitatively differs from standard network equilibria. In particular, we show how standard predictive objectives can drive networks toward consensus even under conditions where classical opinion-dynamics models would suggest disagreement. This emerges because predictive systems dynamically adapt to changing opinions, and learning objectives create spillover effects among individuals beyond the topology of the network.

More formally, study the Friedkin Johnsen model, combined with repeated self-fulfilling predictions biasing individuals' initial opinions. Under this model we first analyze the homogenizing force of performativity for connected networks in the case of perfect prediction. Here we show that opinion can converge to consensus in the limit of high platform susceptibility, even if individuals remain stubborn during peer interactions. This happens because predictions for each individual closely track their opinions during peer interactions. As a result, disagreement reduces with every retraining loop. Even if opinion dynamics alone lead to disagreement. Then, we demonstrate how partial access to data can lead to consensus under even weaker conditions on the network. In particular, we prove that it is sufficient for information to flow in the observed subgroup for a consensus to emerge across the entire population, again, in the limit of high platform susceptibility. The reason is that individuals can indirectly be impacted by the opinions of their peers through model predictions, even if they are topologically isolated. In that way the platform adds an overlaying structure to the social network, determined by the learning objective. We offer a non-parametric lower bound on this performativity-mediated spillover effect unique to predictive loops. Finally, through simulations with parametric models we demonstrate the generality of our claims.

Taken together, our work builds on concepts from network science and performative prediction, and contributes to both. On one hand, it contributes a new perspective to performative prediction by studying stability in a setting where the distribution shift is microfounded by classical model of opinion dynamics and described by a structured population. On the other hand, it contributes to the literature on social networks by studying platform predictions as an active participant in the opinion formation process, and demonstrates performativity as an important, yet so far neglected, qualifying factor in social networks. We believe our work can offer an interesting perspective to the program of the workshop and we would be excited to contribute to the discussion in San Diego.