Understanding filter bubbles – 5’ pitch

# Paradox

Social media offers communication of vast **diversity of information and perspectives**

Social media tend to **recommend us to connect to people with similar opinions** (more interaction, online activity means more interest from advertisers)

Recommendations or filtering **causes creation of filter bubbles** (a.k.a. echo chambers)

* We get more of the same (dumbing down)
* Increasingly, expressed opions get polarized w.r.t. other filter bubbles

# Filter bubbles

A social network can be modeled as a undirected, weighted graph. The nodes represent users and edges to other nodes represent neighbour users influencing each other’s opinion.

**Friedkin-Johnsen (FJ) dynamics** model applies to any social network that can be modeled as an undirected, weighted graph. This model each node is linked to an innate opinion which doesn’t change, and an expressed opinion which is. The FJ is known to converge to an equilibrium set of opinions.

Important concepts in filter bubbles are:

* **Disagreement**: measures misalignment between each node’s opinion and its neighbors.
* **Polarization**: variance of a given set of opinions.
* **Internal conflict**: measures misalignment between each node’s expressed and innate opinion.

Using Friedkin-Johnsen’s dynamics model extended with network administrator actor on Twitter and Reddit datasets, we see:

* Influence of a network administrator (recommender) on polarization by applying small adjustments to user interactions (recommendations or filtered content)
* When reducing disagreement among users, polarization increases.

# Fragile consensus in social network graphs

The equilibirum polarization of a SBM social network plotted as a function of nq (n: 2n nodes, q: probability of out-of-group edge), i.e. the average number of “out-of-group” edges in the network per node. Polarization falls rapidly with nq, leading to a state of potentially fragile consensus, where removing a small number of edges from a network can vastly increase polarization.

# Remedy

We show and proof a remedy to polarization by applying regularization to network administrator dynamics for the Twitter and Reddit datasets. Polarization if vastly reduced where disagreement only slightly increases compared to non-regularized network administrator dynamics.