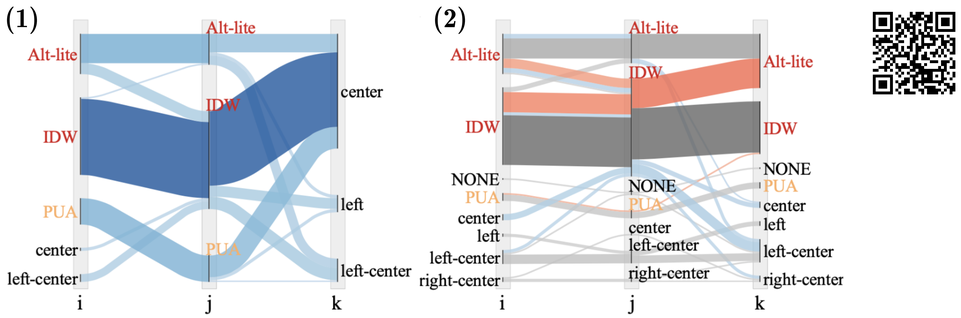
Corinna Coupette
Max Planck Institute for Informatics

Stefan Neumann Aristides Gionis
KTH Royal Institute of Technology

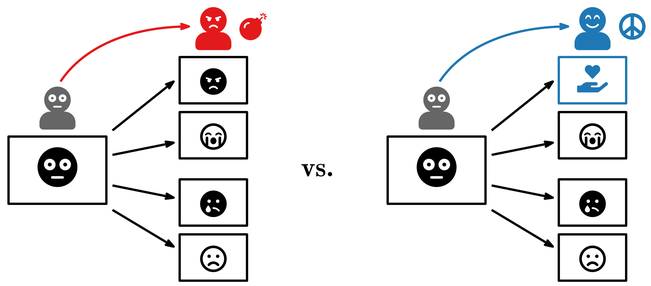
Reducing Exposure to Harmful Content via Graph Rewiring



Edges rewired by GAMINE without (1) or with strict (2) quality constraints.

Motivation

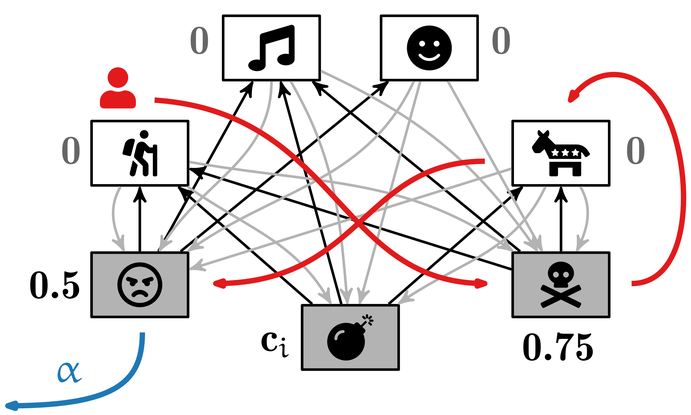
Radicalization Risks of Recommendation Algorithms



How can we mitigate the risks of recommendation algorithms by making small changes to the structure of the recommendation graph?

Harm Exposure Model

Exposure to Harm: Cost of Random Walks on the Graph



Problem Statement

Edge Rewiring: Replacing (i, j) with (i, k)

Other Building Blocks

- Random-walk transition matrix $\mathbf{P}$
- Fundamental matrix $\mathbf{F} = (\mathbf{I} - \mathbf{P})^{-1}$
- G_r: G after r rewirings
- $\mathbf{e}_i^\top \mathbf{F}\mathbf{c}$: Expected exposure starting at i, with $\mathbf{c}$: node-cost vector
- $f(G) = \mathbf{1}^\top \mathbf{F}\mathbf{c}$: Expected total exposure

r-Rewiring Exposure Minimization (REM)

$$\min f(G_r) \Leftrightarrow \max f_\Delta(G, G_r) = f(G) - f(G_r)$$

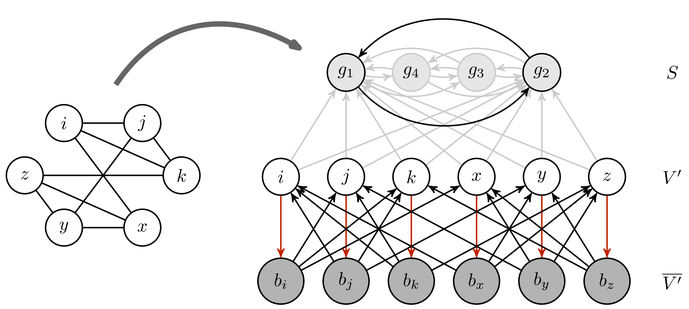
Variant with quality constraints on recommendations:
q-relevant r-Rewiring Exposure Minimization (QREM)
Requires *relevance function* θ (e.g., NDCG) and threshold q

Hardness and Approximability

Greedy $(1 - 1/e)$-APX: Conditional Submodularity

- $S = \{i \in V \mid \mathbf{e}_i^\top \mathbf{F}\mathbf{c} = 0\}$: Set of *safe* nodes
- Λ^+: Maximum out-degree of an *unsafe* node
- $|S| \geqslant \Lambda^+ \Rightarrow$ REM is submodular $\Rightarrow$ greedy $(1 - 1/e)$-APX

NP-Hardness: Reduction from MVC for 3-Regular Graphs



The GAMINE Algorithm

Efficient Implementation

- **Naïve Approach:** $O(rn^2(n+m))$
  Bottleneck: Matrix inversion
- **Forgoing Matrix Inversion:** $O(r\kappa n(n+m))$
  Approximate inverse via κ power iterations
  New bottleneck: Number of candidate rewirings
- **Reducing Candidate Rewirings:** $O(r\kappa(\Delta^+ n + m))$
  REM: Only consider $\Delta^+ + 2$ most promising targets, where Δ^+ is the maximum out-degree in G

Can find rewiring maximizing $\sigma\tau = (\mathbf{1}^\top \mathbf{F}\mathbf{u})(\mathbf{v}^\top \mathbf{F}\mathbf{c})$
Can no longer compute $\rho = 1 + \mathbf{v}^\top \mathbf{F}\mathbf{u}$, but…

- Correlation between $\sigma\tau$ and $\sigma\tau/\rho$ almost perfect
- $\sigma\tau > \sigma\tau'$ *almost* always implies $\sigma\tau/\rho > \sigma\tau'/\rho'$

Linear under realistic assumptions on the input

Experimental Evaluation

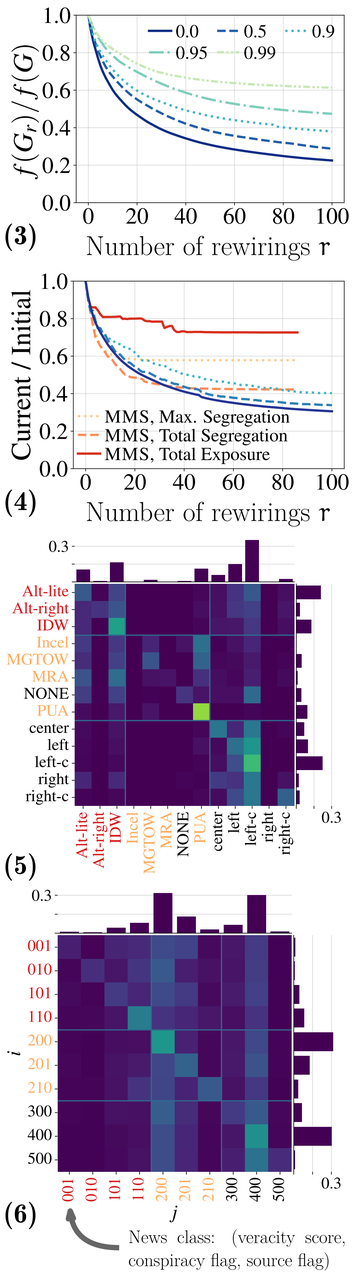
Setup

- Synthetic data & real data (YouTube & NELA-GT)
- 4 different cost functions for real data
- 5 quality thresholds, 3 absorption probabilities, …
- 4 baselines & 1 external competitor (MMS)

With just 100 rewirings, GAMINE can reduce the exposure to harm by 50% while reducing recommendation quality by at most 5% (3).

Selected Observations

- Rewiring to harmful nodes may be necessary (1,2).
- GAMINE outperforms its competitor MMS on the YouTube data (4).
- The NELA-GT data is intrinsically harder than the YouTube data due to its edge structure (5,6).



European Research Council
Established by the European Commission
ACADEMY OF FINLAND
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM