Reducing Exposure to Harmful Content via Graph Rewiring

Corinna Coupette
MPI for Informatics
Germany

Stefan Neumann
KTH Royal Institute of Technology
Sweden

Aristides Gionis
KTH Royal Institute of Technology
Sweden

ABSTRACT

Most media content consumed today is provided by digital platforms that aggregate input from diverse sources, where access to information is mediated by recommendation algorithms. One principal challenge in this context is dealing with content that is considered harmful. Striking a balance between competing stakeholder interests, rather than block harmful content altogether, one approach is to minimize the exposure to such content that is induced specifically by algorithmic recommendations. Hence, modeling media items and recommendations as a directed graph, we study the problem of reducing the exposure to harmful content via edge rewiring. We formalize this problem using absorbing random walks, and prove that it is NP-hard and NP-hard to approximate to within an additive error, while under realistic assumptions, the greedy method yields a $(1 - 1/e)$-approximation. Thus, we introduce Gamine, a fast greedy algorithm that can reduce the exposure to harmful content with or without quality constraints on recommendations. By performing just 100 rewirings on YouTube graphs with several hundred thousand edges, Gamine reduces the initial exposure by 50%, while ensuring that its recommendations are at most 5% less relevant than the original recommendations. Through extensive experiments on synthetic data and real-world data from video recommendation and news feed applications, we confirm the effectiveness, robustness, and efficiency of Gamine in practice.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; Web applications; • **Theory of computation** → **Random walks and Markov chains**; • **Mathematics of computing** → *Graph algorithms.*

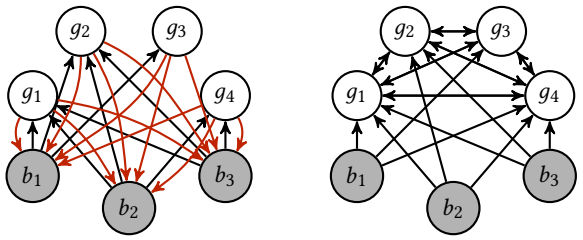
KEYWORDS

graph rewiring, random walks, recommendation graphs

ACM Reference Format:
Corinna Coupette, Stefan Neumann, and Aristides Gionis. 2023. Reducing Exposure to Harmful Content via Graph Rewiring. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599489

1 INTRODUCTION

Recommendation algorithms mediate access to content on digital platforms, and as such, they critically influence how individuals and societies perceive the world and form their opinions [12, 21, 36, 42, 44]. In recent years, platforms have come under increasing scrutiny from researchers and regulators alike due to concerns and evidence that their recommendation algorithms create filter bubbles [6, 26, 28, 45] and fuel radicalization [19, 27, 39, 41, 49]. One of the main challenges in this context is dealing with content that is considered harmful [4, 7, 50]. To address this challenge while balancing the interests of creators, users, and platforms, rather than block harmful content, one approach is to minimize the exposure to such content that is induced by algorithmic recommendations.



(a) Minimizing *segregation* **(b) Minimizing *exposure***

Figure 1: 3-out-regular directed graphs with four *good* nodes (white) and three *bad* nodes (gray). Edges running from good to bad nodes are drawn in red. The left graph minimizes the *segregation* objective from Fabbri et al. [10], but random walks oscillate between good nodes and bad nodes. In contrast, only the right graph minimizes our *exposure* objective.

In this paper, we study the problem of reducing the exposure to harmful content via *edge rewiring*, i.e., replacing certain recommendations by others. This problem was recently introduced by Fabbri et al. [10], who proposed to address it by modeling harmfulness as a *binary* node label and minimizing the *maximum segregation*, defined as the largest expected number of steps of a random walk starting at a harmful node until it visits a benign node. However, while Fabbri et al. [10] posed a theoretically interesting and practically important problem, their approach has some crucial limitations.

First, treating harmfulness as dichotomous fails to capture the complexity of real-world harmfulness assessments. Second, the segregation objective ignores completely all random-walk continuations that return to harmful content after the first visit to a benign node, but *benign nodes do not act as absorbing states* in practice. The consequences are illustrated in Fig. 1a, where the segregation objective judges that the graph provides minimal exposure to harmful content (the hitting time from any harmful node to a benign node is 1), while long random walks, which model user behavior more realistically, oscillate between harmful and benign content.

In this paper, we remedy the above-mentioned limitations. First, we more nuancedly model harmfulness as *real-valued* node costs. Second, we propose a novel minimization objective, the *expected total exposure*, defined as the sum of the costs of absorbing random walks starting at any node. Notably, in our model, no node is an absorbing state, but any node can lead to absorption, which represents more faithfully how users cease to interact with a platform. Our

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08.
https://doi.org/10.1145/3580305.3599489

exposure objective truly minimizes the exposure to harmful content. For example, it correctly identifies the graph in Fig. 1b as significantly less harmful than that in Fig. 1a, while for the segregation objective by Fabbri et al. [10], the two graphs are indistinguishable.

On the algorithmic side, we show that although minimizing the expected total exposure is NP-hard and NP-hard to approximate to within an additive error, its maximization version is equivalent to a submodular maximization problem under the assumption that the input graph contains a small number of *safe* nodes, i.e., nodes that cannot reach nodes with non-zero costs. If these safe nodes are present—which holds in 80% of the real-world graphs used in our experiments—the greedy method yields a $(1 - 1/e)$-approximation. Based on our theoretical insights, we introduce Gamine, a fast greedy algorithm for reducing exposure to harmful content via edge rewiring. Gamine leverages provable strategies for pruning unpromising rewiring candidates, and it works both with and without quality constraints on recommendations. With just 100 rewirings on YouTube graphs containing hundred thousands of edges, Gamine reduces the exposure by 50%, while ensuring that its recommendations are at least 95% as relevant as the originals.

In the following, we introduce our problems, REM and QREM (Section 2), and analyze them theoretically (Section 3). Building on our theoretical insights, we develop Gamine as an efficient greedy algorithm for tackling our problems (Section 4). Having discussed related work (Section 5), we demonstrate the performance of Gamine through extensive experiments (Section 6) before concluding with a discussion (Section 7). All code, datasets, and results are publicly available,[1] and we provide further materials in Appendices A to F.

2 PROBLEMS

We consider a directed graph $G = (V, E)$ of content items (V) and what-to-consume-next recommendations (E), with $n = |V|$ nodes and $m = |E|$ edges. Since we can typically make a fixed number of recommendations for a given content item, such *recommendation graphs* are often d-out-regular, i.e., all nodes have $d = m/n$ out-neighbors, but we do not restrict ourselves to this setting. Rather, each node i has an out-degree $\delta^+(i) = |\Gamma^+(i)|$, where $\Gamma^+(i)$ is the set of out-neighbors of i, and a cost $c_i \in [0, 1]$, which quantifies the harmfulness of content item i, ranging from 0 (not harmful at all) to 1 (maximally harmful). For convenience, we define $\Delta^+ = \max\{\delta^+(i) \mid i \in V\}$ and collect all costs into a vector $\mathbf{c} \in [0, 1]^n$. We model user behavior as a random-walk process on the recommendation graph G. Each edge (i, j) in the recommendation graph is associated with a transition probability p_{ij} such that $\sum_{j \in \Gamma^+(i)} p_{ij} = 1 - \alpha_i$, where α_i is the absorption probability of a random walk at node i (i.e., the probability that the walk ends at i). Intuitively, one can interpret α_i as the probability that a user stops using the service after consuming content i. For simplicity, we assume $\alpha_i = \alpha \in (0, 1]$ for all $i \in V$. Thus, we can represent the random-walk process on G by the transition matrix $\mathbf{P} \in [0, 1 - \alpha]^{n \times n}$, where

$$\mathbf{P}[i, j] = \begin{cases} p_{ij} & \text{if } (i, j) \in E\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{1}$$

[1] 10.5281/zenodo.7936816

This is an absorbing Markov chain, and the expected number of visits from a node i to a node j before absorption is given by the entry (i, j) of the *fundamental matrix* $\mathbf{F} \in \mathbb{R}_{\geq 0}^{n \times n}$, defined as

$$\mathbf{F} = \sum_{i=0}^{\infty} \mathbf{P}^i = (\mathbf{I} - \mathbf{P})^{-1}\,, \tag{2}$$

where $\mathbf{I}$ is the $n \times n$-dimensional identity matrix, and the series converges since $\|\mathbf{P}\|_\infty = \max_i \sum_{j=0}^{n} \mathbf{P}[i, j] = 1 - \alpha < 1$. Denoting the i-th unit vector as $\mathbf{e}_i$, observe that the row vector $\mathbf{e}_i^T\mathbf{F}$ gives the expected number of visits, before absorption, from i to any node, and the column vector $\mathbf{F}\mathbf{e}_i$ gives the expected number of visits from any node to i. Hence, $\mathbf{e}_i^T\mathbf{F}\mathbf{c} = \sum_{j \in V} \mathbf{F}[i, j]\mathbf{c}_j$ gives the expected exposure to harmful content of users starting their random walk at node i, referred to as the *exposure* of i. The *expected total exposure* to harm in the graph G, then, is given by the non-negative function

$$f(G) = \mathbf{1}^T\mathbf{F}\mathbf{c}\,, \tag{3}$$

where $\mathbf{1}$ is the vector with each entry equal to 1.

We would like to *minimize* the exposure function given in Eq. (3) by making r edits to the graph G, i.e., we seek an effective *post-processing strategy* for harm reduction. In line with our motivating application, we restrict edits to *edge rewirings* denoted as (i, j, k), in which we replace an edge $(i, j) \in E$ by an edge $(i, k) \notin E$ with $i \neq k$, setting $p_{ik} = p_{ij}$ (other edits are discussed in Appendix B). Seeking edge rewirings to minimize the expected total exposure yields the following problem definition.

Problem 1 (r-rewiring exposure minimization [REM])**.** *Given a graph G, its random-walk transition matrix $\mathbf{P}$, a node cost vector $\mathbf{c}$, and a budget r, minimize $f(G_r)$, where G_r is G after r rewirings.*

Equivalently, we can *maximize* the *reduction* in the expected total exposure to harmful content,

$$f_\Delta(G, G_r) = f(G) - f(G_r)\,. \tag{4}$$

Note that while any set of rewirings minimizing $f(G_r)$ also maximizes $f_\Delta(G, G_r)$, the approximabilities of f and f_Δ can differ widely.

As Problem 1 does not impose any constraints on the rewiring operations, the optimal solution might contain rewirings (i, j, k) such that node k is unrelated to i. To guarantee high-quality recommendations, we need additional relevance information, which we assume to be given as a *relevance matrix* $\mathbf{R} \in \mathbb{R}_{\geq 0}^{n \times n}$, where $\mathbf{R}[i, j]$ denotes the relevance of node j in the context of node i. Given such relevance information, and assuming that the out-neighbors of a node i are ordered as $\mathbf{r}_i \in V^{\delta^+(i)}$, we can define a *relevance function* θ with range $[0, 1]$ to judge the quality of the recommendation sequence at node i, depending on the relevance and ordering of recommended nodes, and demand that any rewiring retain $\theta(\mathbf{r}_i) \geq q$ for all $i \in V$ and some *quality threshold* $q \in [0, 1]$. One potential choice for θ is the normalized discounted cumulative gain (nDCG), a popular ranking quality measure, which we use in our experiments and define in Appendix D.1. Introducing θ allows us to consider a variant of REM with relevance constraints.

Problem 2 (q-relevant r-rewiring exposure minimization [QREM])**.** *Given a graph G, its random-walk transition matrix $\mathbf{P}$, a node cost vector $\mathbf{c}$, a budget r, a relevance matrix $\mathbf{R}$, a relevance function θ, and a quality threshold q, minimize $f(G_r)$ under the condition that $\theta(\mathbf{r}_i) \geq q$ for all $i \in V$.*

For $q = 0$, QREM is equivalent to REM. Collecting our notation in Appendix Table 3, we now seek to address both problems.

3 THEORY

To start with, we establish some theoretical properties of our problems, the functions f and f_Δ, and potential solution approaches.

Hardness. We begin by proving that REM (and hence, also QREM) is an NP-hard problem.

Theorem 1 (NP-Hardness of REM). *The r-rewiring exposure minimization problem is* NP*-hard, even on 3-out-regular input graphs with binary costs $\mathbf{c} \in \{0,1\}^n$.*

Proof. We obtain this result by reduction from minimum vertex cover for cubic, i.e., 3-regular graphs (MVC-3), which is known to be NP-hard [16]. A full, illustrated proof is given in Appendix A.1. □

Next, we further show that REM is hard to approximate under the Unique Games Conjecture (UGC) [24], an influential conjecture in hardness-of-approximation theory.

Theorem 2. *Assuming the UGC, REM is hard to approximate to within an additive error of both $\Theta(n)$ and $\Theta(r)$.*

Proof. We obtain this result via the hardness of approximation of MVC under the UGC. A full proof is given in Appendix A.2. □

Both Theorem 1 and Theorem 2 extend from f to f_Δ (Eq. (9)).

Approximability. Although we cannot approximate f directly, we *can* approximate f_Δ *with guarantees* under mild assumptions, detailed below. To formulate this result and its assumptions, we start by calling a node *safe* if $\mathbf{e}_i^T \mathbf{F}\mathbf{c} = 0$, i.e., no node j with $c_j > 0$ is reachable from i, and *unsafe* otherwise. Note that the existence of a safe node in a graph G containing at least one unsafe node (i.e., $c_i > 0$ for some $i \in V$) implies that G is not strongly connected. The node safety property partitions V into two sets of safe resp. unsafe nodes, $S = \{i \in V \mid \mathbf{e}_i^T \mathbf{F}\mathbf{c} = 0\}$ and $U = \{i \in V \mid \mathbf{e}_i^T \mathbf{F}\mathbf{c} > 0\}$, and E into four sets, E_{SS}, E_{SU}, E_{US}, and E_{UU}, where $E_{AB} = \{(i,j) \in E \mid i \in A, j \in B\}$, and $E_{SU} = \emptyset$ by construction. Further, observe that if $S \neq \emptyset$, then f is minimized, and f_Δ is maximized, once $E_{UU} = \emptyset$. This allows us to state the following result.

Lemma 1. *If there exists a safe node in G and we allow multi-edges, maximizing f_Δ is equivalent to maximizing a monotone, submodular set function over E_{UU}.*

Proof. Leveraging the terminology introduced above, we obtain this result by applying the definitions of monotonicity and submodularity. A full proof is given in Appendix A.3. □

Our motivating application, however, ideally prevents multi-edges. To get a similar result without multi-edges, denote by $\Lambda^+ = \max\{\delta^+(i) \mid i \in U\}$ the maximum out-degree of any *unsafe* node in G, and assume that $|S| \geq \Lambda^+$. Now, we obtain the following.

Theorem 3. *If $|S| \geq \Lambda^+$, then maximizing f_Δ is equivalent to maximizing a monotone and submodular set function over E_{UU}.*

Proof. Following the reasoning provided for Lemma 1, with the modification that we need $|S| \geq \Lambda^+$ to ensure that safe targets are always available for rewiring without creating multi-edges. □

Observe that the larger the number of zero-cost nodes, the smaller the number of edges, or the more homophilous the linking, the higher the probability that safe nodes exist in a graph. Notably, the precondition of Theorem 3 holds for the graph constructed to prove Theorem 1 (Appendix A.1, Fig. 10) as well as for most of the real-world graphs used in our experiments (Appendix E, Fig. 17). However, Theorem 3 only applies to the maximization version of REM (Eq. (9)) and not to the maximization version of QREM, since in the quality-constrained setting, some safe nodes might not be available as rewiring targets for edges emanating from unsafe nodes. Still, for the maximization version of REM, due to Theorem 3, using a greedy approach to optimize f_Δ provides an approximation guarantee with respect to the optimal solution [34].

Corollary 1. *If the precondition of Theorem 3 holds, then the greedy algorithm, which always picks the rewiring (i,j,k) that maximizes $f_\Delta(G, G_1)$ for the current G, yields a $(1 - 1/e)$-approximation for f_Δ.*

Note that Corollary 1 only applies to the *maximization* version of REM, not to its *minimization* version, as supermodular minimization is less well-behaved than submodular maximization [22, 52].

Greedy Rewiring. Given the quality assurance of a greedy approach at least for REM, we seek to design an efficient greedy algorithm to tackle both REM and QREM. To this end, we analyze the mechanics of individual rewirings to understand how we can identify and perform greedily optimal rewirings efficiently. As each greedy step constitutes a rank-one update of the transition matrix $\mathbf{P}$, we can express the new transition matrix $\mathbf{P}'$ as

$$\mathbf{P}' = \mathbf{P} + \mathbf{u}(-\mathbf{v})^T \,, \tag{5}$$

where $\mathbf{u} = p_{ij}\mathbf{e}_i$ and $\mathbf{v} = \mathbf{e}_j - \mathbf{e}_k$, and we omit the dependence on i, j, and k for notational conciseness. This corresponds to a rank-one update of $\mathbf{F}$, such that we obtain the new fundamental matrix $\mathbf{F}'$ as

$$\mathbf{F}' = (\mathbf{I} - (\mathbf{P} + \mathbf{u}(-\mathbf{v})^T))^{-1} = (\mathbf{I} - \mathbf{P} + \mathbf{u}\mathbf{v}^T)^{-1} \,. \tag{6}$$

The rank-one update allows us to use the Sherman-Morrison formula [43] to compute the updated fundamental matrix as

$$\mathbf{F}' = \mathbf{F} - \frac{\mathbf{F}\mathbf{u}\mathbf{v}^T\mathbf{F}}{1 + \mathbf{v}^T\mathbf{F}\mathbf{u}} \,. \tag{7}$$

The mechanics of an individual edge rewiring are summarized in Table 1. They will help us *perform* greedy updates efficiently.

To also *identify* greedily optimal rewirings efficiently, leveraging Eq. (7), we assess the impact of a rewiring on the value of our objective function, which will help us prune weak rewiring candidates. For a rewiring (i,j,k) represented by $\mathbf{u}$ and $\mathbf{v}$, the value of the exposure function f for the new graph G' is

$$\begin{aligned} f(G') &= \mathbf{1}^T\mathbf{F}'\mathbf{c} = \mathbf{1}^T\left(\mathbf{F} - \frac{\mathbf{F}\mathbf{u}\mathbf{v}^T\mathbf{F}}{1 + \mathbf{v}^T\mathbf{F}\mathbf{u}}\right)\mathbf{c} = \mathbf{1}^T\mathbf{F}\mathbf{c} - \mathbf{1}^T\left(\frac{\mathbf{F}\mathbf{u}\mathbf{v}^T\mathbf{F}}{1 + \mathbf{v}^T\mathbf{F}\mathbf{u}}\right)\mathbf{c} \\ &= f(G) - \frac{(\mathbf{1}^T\mathbf{F}\mathbf{u})(\mathbf{v}^T\mathbf{F}\mathbf{c})}{1 + \mathbf{v}^T\mathbf{F}\mathbf{u}} = f(G) - \frac{\sigma\tau}{\rho} = f(G) - \Delta \,, \end{aligned} \tag{8}$$

with $\sigma = \mathbf{1}^T\mathbf{F}\mathbf{u}$, $\tau = \mathbf{v}^T\mathbf{F}\mathbf{c}$, $\rho = 1 + \mathbf{v}^T\mathbf{F}\mathbf{u}$, and

$$\Delta = f_\Delta(G, G') = \frac{\sigma\tau}{\rho} = \frac{(\mathbf{1}^T\mathbf{F}\mathbf{u})(\mathbf{v}^T\mathbf{F}\mathbf{c})}{1 + \mathbf{v}^T\mathbf{F}\mathbf{u}} \,. \tag{9}$$

The interpretation of the above quantities is as follows: σ is the p_{ij}-scaled i-th column sum of $\mathbf{F}$ (expected number of visits to i), τ is

Table 1: Summary of an edge rewiring (i, j, k) in a graph $G = (V, E)$ with random-walk transition matrix $\mathbf{P}$ and fundamental matrix $\mathbf{F} = (\mathbf{I} - \mathbf{P})^{-1}$.

$G' = (V, E')$, for $E' = (E \setminus \{(i,j)\}) \cup \{(i,k)\}$, $(i,j) \in E$, $(i,k) \notin E$
$\mathbf{P}'[x,y] = \begin{cases} 0 & \text{if } x = i \text{ and } y = j\,, \\ \mathbf{P}[i,j] & \text{if } x = i \text{ and } y = k\,, \\ \mathbf{P}[x,y] & \text{otherwise}\,. \end{cases}$
$\mathbf{F}' = \mathbf{F} - \frac{\mathbf{F}\mathbf{u}\mathbf{v}^T\mathbf{F}}{1+\mathbf{v}^T\mathbf{F}\mathbf{u}}$, with $\mathbf{u} = p_{ij}\mathbf{e}_i$, $\mathbf{v} = \mathbf{e}_j - \mathbf{e}_k$, cf. Eq. (7)

the cost-scaled sum of the differences between the j-th row and the k-th row of $\mathbf{F}$ (expected number of visits from j resp. k), and ρ is a normalization factor scaling the update by 1 plus the p_{ij}-scaled difference in the expected number of visits from j to i and from k to i, ensuring that $\mathbf{F}'\mathbf{1} = \mathbf{F}\mathbf{1}$. Scrutinizing Eq. (9), we observe:

Lemma 2. *For a rewiring (i, j, k) represented by $\mathbf{u}$ and $\mathbf{v}$, (i) ρ is always positive, (ii) σ is always positive, and (iii) τ can have any sign.*

Proof. We obtain this result by analyzing the definitions of ρ, σ, and τ. The full proof is given in Appendix A.4. □

To express when we can safely prune rewiring candidates, we call a rewiring (i, j, k) *greedily permissible* if $\Delta > 0$, i.e., if it reduces our objective, and *greedily optimal* if it maximizes Δ. For QREM, we further call a rewiring (i, j, k) *greedily q-permissible* if it ensures that $\theta(\mathbf{r}_i) \geq q$ under the given relevance function θ. With this terminology, we can confirm our intuition about rewirings as a corollary of Eqs. (8) and (9), combined with Lemma 2.

Corollary 2. *A rewiring (i, j, k) is greedily permissible if and only if $\tau > 0$, i.e., if j is more exposed to harm than k.*

For the greedily optimal rewiring, that is, to maximize Δ, we would like $\sigma\tau$ to be as *large* as possible, and ρ to be as *small* as possible. Inspecting Eq. (9), we find that to accomplish this objective, it helps if (in expectation) i is visited more often (from σ), j is more exposed and k is less exposed to harm (from τ), and i is harder to reach from j and easier to reach from k (from ρ).

In the next section, we leverage these insights to guide our efficient implementation of the greedy method for REM and QREM.

4 ALGORITHM

In the previous section, we identified useful structure in the fundamental matrix $\mathbf{F}$, the exposure function f, and our maximization objective f_Δ. Now, we leverage this structure to design an efficient greedy algorithm for REM and QREM. We develop this algorithm in three steps, focusing on REM in the first two steps, and integrating the capability to handle QREM in the third step.

Naïve implementation. Given a graph G, its transition matrix $\mathbf{P}$, a cost vector $\mathbf{c}$, and a budget r, a naïve greedy implementation for REM computes the fundamental matrix and gradually fills up an initially empty set of rewirings by performing r greedy steps before returning the selected rewirings (Appendix C, Algorithm 3). In each greedy step, we identify the triple (i, j, k) that maximizes Eq. (9) by going through all edges $(i, j) \in E$ and computing Δ for rewirings to all potential targets k. We then update E, $\mathbf{P}$, and $\mathbf{F}$ to reflect a rewiring replacing (i, j) by (i, k) (cf. Table 1), and add the triple (i, j, k) to our set of rewirings. Computing the fundamental matrix naïvely takes time $O(n^3)$, computing Δ takes time $O(n)$ and is done $O(mn)$ times, and updating $\mathbf{F}$ takes time $O(n^2)$. Hence, we arrive at a time complexity of $O(rn^2(n+m))$. But we can do better.

Forgoing matrix inversion. When identifying the greedy rewiring, we never need access to $\mathbf{F}$ directly. Rather, in Eq. (9), we work with $\mathbf{1}^T\mathbf{F}$, corresponding to the column sums of $\mathbf{F}$, and with $\mathbf{F}\mathbf{c}$, corresponding to the cost-scaled row sums of $\mathbf{F}$. We can approximate both via power iteration:

$$\mathbf{1}^T\mathbf{F} = \mathbf{1}^T \sum_{i=0}^{\infty} \mathbf{P}^i = \mathbf{1}^T + \mathbf{1}^T\mathbf{P} + (\mathbf{1}^T\mathbf{P})\mathbf{P} + ((\mathbf{1}^T\mathbf{P})\mathbf{P})\mathbf{P} + \dots \tag{10}$$

$$\mathbf{F}\mathbf{c} = \left(\sum_{i=0}^{\infty} \mathbf{P}^i\right)\mathbf{c} = \mathbf{c} + \mathbf{P}\mathbf{c} + \mathbf{P}(\mathbf{P}\mathbf{c}) + \mathbf{P}(\mathbf{P}(\mathbf{P}\mathbf{c})) + \dots \tag{11}$$

For each term in these sums, we need to perform $O(m)$ multiplications, such that we can compute $\mathbf{1}^T\mathbf{F}$ and $\mathbf{F}\mathbf{c}$ in time $O(\kappa m)$, where κ is the number of power iterations. This allows us to compute $\mathbf{1}^T\mathbf{F}\mathbf{u}$ for all $(i, j) \in E$ in time $O(m)$ and $\mathbf{v}^T\mathbf{F}\mathbf{c}$ for all $j \neq k \in V$ in time $O(n^2)$. To compute Δ in time $O(1)$, as $\mathbf{F}$ is now unknown, we need to compute $\mathbf{F}\mathbf{u}$ for all $(i, j) \in E$ via power iteration, which is doable in time $O(\kappa n^2)$. This changes the running time from $O(rn^2(n+m))$ to $O(r\kappa n(n+m))$ (Appendix C, Algorithm 4). But we can do better.

Reducing the number of candidate rewirings. Observe that to further improve the time complexity of our algorithm, we need to reduce the number of rewiring candidates considered. To this end, note that the quantity τ is maximized for the nodes j and k with the largest difference in cost-scaled row sums. How exactly we leverage this fact depends on our problem.

If we solve REM, instead of considering all possible rewiring targets, we focus on the $\Delta^+ + 2$ candidate targets K with the smallest exposure, which we can identify in time $O(n)$ without sorting $\mathbf{F}\mathbf{c}$. This ensures that for each $(i, j) \in E$, there is at least one $k \in K$ such that $k \neq i$ and $k \neq j$, which ascertains that despite restricting to K, for each $i \in V$, we still consider the rewiring (i, j, k) maximizing τ. With this modification, we reduce the number of candidate targets from $O(n)$ to $O(\Delta^+)$ and the time to compute all relevant $\mathbf{v}^T\mathbf{F}\mathbf{c}$ values from $O(n^2)$ to $O(\Delta^+ n)$. To obtain a subquadratic complexity, however, we still need to eliminate the computation of $\mathbf{F}\mathbf{u}$ for all $(i, j) \in E$. This also means that we can no longer afford to compute ρ for each of the now $O(m\Delta^+)$ rewiring candidates under consideration, as this can only be done in constant time if $\mathbf{F}\mathbf{u}$ is already precomputed for the relevant edge (i, j). However, ρ is driven by the difference between two *entries* of $\mathbf{F}$, whereas τ is driven by the difference between two *row sums* of $\mathbf{F}$, and σ is driven by a single *column sum* of $\mathbf{F}$. Thus, although $\sigma\tau > \sigma\tau'$ does not generally imply $\sigma\tau/\rho > \sigma\tau'/\rho'$, the variation in $\sigma\tau$ is typically much larger than that in ρ, and large $\sigma\tau$ values mostly dominate small values of ρ. Consequently, as demonstrated in Appendix F.3, the correlation between $\hat{\Delta} = \Delta\rho = \sigma\tau$ and $\Delta = \sigma\tau/\rho$ is almost perfect. Thus, instead of Δ, we opt to compute $\hat{\Delta}$ as a heuristic, and we further hedge against small fluctuations without increasing the time complexity of our algorithm by computing Δ for the rewirings associated with the $O(1)$ *largest* values of $\hat{\Delta}$, rather than selecting

the rewiring with the *best* $\hat{\Delta}$ value directly. Using $\hat{\Delta}$ instead of Δ, we obtain a running time of $O(r\kappa\Delta^+(n+m))$ when solving REM.

When solving QREM, we are given a relevance matrix $\mathbf{R}$, a relevance function θ, and a relevance threshold q as additional inputs. Instead of considering the $\Delta^+ + 2$ nodes K with the smallest exposure as candidate targets for *all* edges, for *each* edge (i, j), we first identify the set of rewiring candidates (i, j, k) such that (i, j, k) is q-permissible, i.e., $\theta(\mathbf{r}_i) \geq q$ after replacing (i, j) by (i, k), and then select the node k_{ij} with the smallest exposure to construct our most promising rewiring candidate (i, j, k_{ij}) for edge (i, j). This ensures that we can still identify the rewiring (i, j, k) that maximizes $\sigma\tau$ *and* satisfies our quality constraints, and it leaves us to consider $O(m)$ rewiring candidates. Again using $\hat{\Delta}$ instead of Δ, we can now solve QREM in time $O(r\kappa\ell gm + h)$, where ℓ is the maximum number of targets k such that (i, j, k) is q-permissible, g is the complexity of evaluating θ, and h is the complexity of determining the initial set Q of q-permissible rewirings.

Thus, we have arrived at our efficient greedy algorithm, called Gamine (Greedy Approximate miNimization of Exposure), whose pseudocode we state as Algorithms 1 and 2 in Appendix C. Gamine solves REM in time $O(r\kappa\Delta^+(n+m))$ and QREM in time $O(r\kappa\ell gm + h)$. In realistic recommendation settings, the graph G is d-out-regular for $d \in O(1)$, such that $\Delta^+ \in O(1)$ and $m = dn \in O(n)$. Further, for QREM, we can expect that θ is evaluable in time $O(1)$, and that only the $O(1)$ nodes most relevant for i will be considered as potential rewiring targets of any edge (i, j), such that $\ell \in O(1)$ and $h \in O(m) = O(n)$. As we can also safely work with a number of power iterations $\kappa \in O(1)$ (Appendix D.3), in realistic settings, Gamine solves both REM and QREM in time $O(rn)$, which, for $r \in O(1)$, is linear in the order of the input graph G.

5 RELATED WORK

Our work methodically relates to research on *graph edits* with distinct goals, such as improving robustness, reducing distances, or increasing centralities [5, 32, 37], and research leveraging *random walks* to rank nodes [30, 35, 48] or recommend links [38, 51]. The agenda of our work, however, aligns most closely with the literature studying harm reduction, bias mitigation, and conflict prevention in graphs. Here, the large body of research on shaping opinions or mitigating negative phenomena in *graphs of user interactions* (especially on social media) [1, 3, 8, 13–15, 33, 46, 47, 53, 54] pursues goals *similar* to ours in graphs capturing *different* digital contexts.

As our research is motivated by recent work demonstrating how recommendations on digital media platforms like YouTube can fuel radicalization [29, 41], the comparatively scarce literature on harm reduction in *graphs of content items* is even more closely related. Our contribution is inspired by Fabbri et al. [10], who study how edge rewiring can reduce *radicalization pathways* in recommendation graphs. Fabbri et al. [10] encode harmfulness in binary node labels, model benign nodes as absorbing states, and aim to minimize the *maximum segregation* of any node, defined as the largest expected length of a random walk starting at a harmful node before it visits a benign node. In contrast, we encode harmfulness in more nuanced, real-valued node attributes, use an absorbing Markov chain model that more naturally reflects user behavior, and aim to minimize the *expected total exposure* to harm in random walks starting at any node. Thus, our work not only eliminates several limitations of the work by Fabbri et al. [10], but it also provides a different perspective on harm mitigation in recommendation graphs.

While Fabbri et al. [10], like us, consider *recommendation graphs*, Haddadan et al. [18] focus on polarization mitigation via *edge insertions*. Their setting was recently reconsidered by Adriaens et al. [2], who tackle the minimization objective directly instead of using the maximization objective as a proxy, providing approximation bounds as well as speed-ups for the standard greedy method. Both Fabbri et al. [10] and the works on edge insertion employ with random-walk objectives that—unlike our exposure function—do not depend on random walks starting from *all* nodes. In our experiments, we compare with the algorithm introduced by Fabbri et al. [10], which we call MMS. We refrain from comparing with edge insertion strategies because they consider a different graph edit operation and are already outperformed by MMS.

6 EXPERIMENTAL EVALUATION

In our experiments, we seek to

(1) establish the impact of modeling choices and input parameters on the performance of Gamine;
(2) demonstrate the effectiveness of Gamine in reducing exposure to harm compared to existing methods and baselines;
(3) ensure that Gamine is scalable in theory *and practice*;
(4) understand what features make reducing exposure to harm easier resp. harder on different datasets; and
(5) derive general guidelines for reducing exposure to harm in recommendation graphs under budget constraints.

Further experimental results are provided in Appendix F.

6.1 Setup

6.1.1 Datasets. To achieve our experimental goals, we work with both synthetic and real-world data, as summarized in Table 2. Below, we briefly introduce these datasets. Further details, including on data generation and preprocessing, are provided in Appendix E.

Synthetic data. As our synthetic data, we generate a total of 288 synthetic graphs of four different sizes using two different edge placement models and various parametrizations. The first model, SU, chooses out-edges *uniformly* at random, similar to a directed Erdős-Rényi model [9]. In contrast, the second model, SH, chooses edges *preferentially* to favor small distances between the costs of the source and the target node, implementing the concept of *homophily* [31]. We use these graphs primarily to analyze the behavior of our objective function, and to understand the impact of using $\hat{\Delta}$ instead of Δ to select the greedily optimal rewiring (Appendix F.3).

Real-world data. We work with real-world data from two domains, video recommendations (YT) and news feeds (NF). For our *video application*, we use the YouTube data by Ribeiro et al. [29, 41], which contains identifiers and "Up Next"-recommendations for videos from selected channels categorized to reflect different degrees and directions of radicalization. For our *news application*, we use subsets of the NELA-GT-2021 dataset [17], which contains 1.8 million news articles published in 2021 from 367 outlets, along with veracity labels from Media Bias/Fact Check. Prior versions of both datasets are used in the experiments reported by Fabbri et al. [10].

Table 2: Overview of the datasets used in our experiments. For each graph G, we report the regular out-degree d, the number of nodes n, and the number of edges m, as well as the range of the expected exposure $f(G)/n$ under our various cost functions, edge wirings, and edge transition probabilities. Datasets with identical statistics are pooled in the same row.

Dataset	d	n	m	$f(G)/n$
SU, SH (2·4·36 graphs)	5	10^i for $i \in \{2,3,4,5\}$	$5{\times}10^i$	[1.291, 15.231]
YT-100K (3 · 6 graphs)	5	40 415	202 075	[0.900, 8.475]
	10		404 150	[0.938, 8.701]
	20		808 300	[0.989, 9.444]
YT-10K (3 · 6 graphs)	5	150 572	752 860	[0.806, 5.785]
	10		1 505 720	[0.883, 7.576]
	20		3 011 440	[0.949, 8.987]
NF-Jan06 (3 · 6 graphs)	5	11 931	59 655	[4.217, 9.533]
	10		119 310	[4.248, 9.567]
	20		238 620	[4.217, 9.533]
NF-Cov19 (3 · 6 graphs)	5	57 447	287 235	[4.609, 11.068]
	10		574 470	[4.392, 10.769]
	20		1 148 940	[4.329, 10.741]
NF-All (3 · 6 graphs)	5	93 455	467 275	[5.565, 11.896]
	10		934 550	[5.315, 11.660]
	20		1 869 100	[5.138, 11.517]

Parametrizations. To comprehensively assess the effect of modeling assumptions regarding the input graph and its associated random-walk process on our measure of exposure as well as on the performance of Gamine and its competitors, we experiment with a variety of parametrizations expressing these assumptions. For all datasets, we distinguish three random-walk absorption probabilities $\alpha \in \{0.05, 0.1, 0.2\}$ and two probability shapes $\chi \in \{\mathbf{U}, \mathbf{S}\}$ over the out-edges of each node (**U**niform and **S**kewed). For our synthetic datasets, we further experiment with three fractions of latently harmful nodes $\beta \in \{0.3, 0.5, 0.7\}$ and two cost functions $c \in \{c_B, c_R\}$, one binary and one real-valued. Lastly, for our real-world datasets, we distinguish three regular out-degrees $d \in \{5, 10, 20\}$, five quality thresholds $q \in \{0.0, 0.5, 0.9, 0.95, 0.99\}$ and four cost functions, two binary (c_B1, c_B2) and two real-valued (c_{R1}, c_{R2}), based on labels provided with the original datasets, as detailed in Appendix E.2.2.

6.1.2 Algorithms. We compare Gamine, our algorithm for REM and QREM, with four baselines (BL1-BL4) and the algorithm by Fabbri et al. [10] for minimizing the maximum segregation, which we call MMS. In all QREM experiments, we use the $O(1)$-computable normalized discounted cumulative gain (nDCG), defined in Appendix D.1 and also used by MMS, as a relevance function θ, and consider the 100 most relevant nodes as potential rewiring targets.

As MMS can only handle binary costs, we transform nonbinary costs c into binary costs c' by thresholding to ensure $c_i \geq \mu \Leftrightarrow c'_i = 1$ for some rounding threshold $\mu \in (0, 1]$ (cf. Appendix D.2). Since MMS requires access to relevance information, we restrict our comparisons with MMS to data where this information is available.

Our baselines BL1-BL4 are ablations of Gamine, such that outperforming them shows how each component of our approach is beneficial. We order the baselines by the competition we expect from them, from no competition at all (BL1) to strong competition (BL4). Intuitively, BL1 does not consider our objective at all, BL2 is a heuristic focusing on the τ component of our objective, BL3 is a heuristic focusing on the σ component of our objective, and BL4 is a heuristic eliminating the iterative element of our approach. BL1–BL3 each run in r rounds, while BL4 runs in one round. In each round, BL1 randomly selects a permissible rewiring via rejection sampling. BL2 selects the rewiring (i, j, k) with the node j maximizing $\mathbf{e}_j^T \mathbf{F}\mathbf{c}$ as its old target, the node i with $j \in \Gamma^+(i)$ maximizing $\mathbf{1}^T \mathbf{F}\mathbf{e}_i$ as its source, and the available node k minimizing $\mathbf{e}_k^T \mathbf{F}\mathbf{c}$ as its new target. BL3 selects the rewiring (i, j, k) with the node i maximizing $\mathbf{1}^T \mathbf{F}\mathbf{e}_i$ as its source, the node j with $j \in \Gamma^+(i)$ maximizing $\mathbf{e}_j^T \mathbf{F}\mathbf{c}$ as its old target, and the available node k minimizing $\mathbf{e}_k^T \mathbf{F}\mathbf{c}$ as its new target. BL4 selects the r rewirings with the largest initial values of $\hat{\Delta}$, while ensuring each edge is rewired at most once.

6.1.3 Implementation and reproducibility. All algorithms, including Gamine, the baselines, and MMS, are implemented in Python 3.10. We run our experiments on a 2.9 GHz 6-Core Intel Core i9 with 32 GB RAM and report wall-clock time. All code, datasets, and results are publicly available,[2] and we provide further reproducibility information in Appendix D.

6.2 Results

6.2.1 Impact of modeling choices. To understand the impact of a particular modeling choice on the performance of Gamine and its competitors, we analyze groups of experimental settings that vary only the parameter of interest while keeping the other parameters constant, focusing on the YT-100K datasets. We primarily report the evolution of the ratio $f(G_r)/f(G) = \big(f(G) - f_\Delta(G, G_r)\big)/f(G)$, which indicates what fraction of the initial expected total exposure is left after r rewirings, and hence is comparable across REM instances with different starting values. Overall, we observe that Gamine robustly reduces the expected total exposure to harm, and that it changes its behavior predictably under parameter variations. Due to space constraints, we defer the results showing this for variations in the regular out-degree d, the random-walk absorption probability α, the probability shape χ, and the cost function c to Appendix F.1.

Impact of quality threshold q. The higher the quality threshold q, the more constrained our rewiring options. Thus, under a given budget r, we expect Gamine to reduce our objective more strongly for smaller q. As illustrated in Fig. 2, our experiments confirm this intuition, and the effect is more pronounced if the out-edge probability distribution is skewed. We further observe that Gamine can guarantee $q = 0.5$ with little performance impact, and it can strongly reduce the exposure to harm even under a strict $q = 0.95$: With just 100 edge rewirings, it reduces the expected total exposure to harm by 50%, while ensuring that its recommendations are at most 5% less relevant than the original recommendations.

6.2.2 Performance comparisons. Having ensured that Gamine robustly and predictably reduces the total exposure across the entire

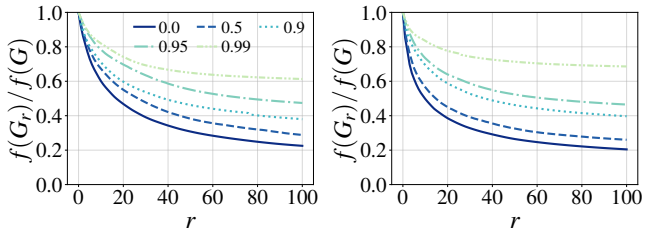
[2] 10.5281/zenodo.7936816

spectrum of modeling choices, we now compare it with its competitors. Overall, we find that GAMINE offers more reliable performance and achieves stronger harm reduction than its contenders.

Comparison with baselines BL1–BL4. First, we compare GAMINE with our four baselines, each representing a different ablation of our algorithm. As depicted in Fig. 3, the general pattern we observe matches our performance expectations (from weak performance of BL1 to strong performance of BL4), but we are struck by the strong performance of BL3 (selecting based on σ), especially in contrast to the weak performance of BL2 (selecting based on τ). This suggests that whereas the most *exposed* node does not necessarily have a highly visited node as an in-neighbor, the most *visited* node tends to have a highly exposed node as an out-neighbor. In other words, for some highly *prominent* videos, the YouTube algorithm problematically appears to recommend highly *harm-inducing* content to watch next. Despite the competitive performance of BL3 and BL4, GAMINE consistently outperforms these baselines, too, and unlike the baselines, it *smoothly* reduces the exposure function. This lends additional support to our reliance on $\sigma\tau$ (rewiring a highly visited i away from a highly exposed j) as an *iteratively* evaluated heuristic.

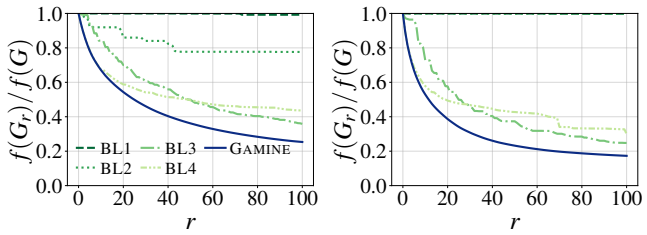
Comparison with MMS. Having established that all components of GAMINE are needed to achieve its performance, we now compare our algorithm with MMS, the method proposed by Fabbri et al. [10]. To this end, we run both GAMINE and MMS using their respective objective functions, i.e., the expected total exposure to harm of random walks starting at any node (*total exposure*, GAMINE) and the maximum expected number of random-walk steps from a harmful node to a benign node (*maximum segregation*, MMS). Reporting their performance under the objectives of *both* algorithms (as well as the *total segregation*, which sums the segregation scores of all harmful nodes) in Fig. 4, we find that under strict quality control ($q \in \{0.9, 0.95, 0.99\}$), GAMINE outperforms MMS on *all* objectives, and MMS stops early as it can no longer reduce its objective function. For $q = 0.5$, MMS outperforms GAMINE on the segregation-based objectives, but GAMINE still outperforms MMS on our exposure-based objective, sometimes at twice the margin (Fig. 4g). Further, while GAMINE delivers consistent and predictable performance that is strong on exposure-based and segregation-based objectives, we observe much less consistency in the performance of MMS. For example, it is counterintuitive that MMS identifies 100 rewirings on the smaller YT-100K data but stops early on the larger YT-10K data. Moreover, MMS delivers the results shown in Fig. 4 under c_{B1}, but it cannot decrease its objective at all on the same data under c_{B2}, which differs from c_{B1} only in that it also assigns harm to anti-feminist content (Appendix E, Table 7). We attribute this brittleness to the reliance on the *maximum*-based *segregation* objective, which, by design, is less robust than our *sum*-based *exposure* objective.

6.2.3 Empirical scalability of GAMINE. In our previous experiments, we found that GAMINE robustly and reliably reduces the expected total exposure to harm. Now, we seek to ascertain that its practical scaling behavior matches our theoretical predictions, i.e., that under realistic assumptions on the input, GAMINE scales linearly in n and m. We are also interested in comparing GAMINE's scalability to that of MMS. To this end, we measure the time taken to compute a single rewiring and report, in Fig. 5, the *average* over ten rewirings



(a) Probability shape χ = U **(b) Probability shape χ = S**

Figure 2: Performance of GAMINE for quality thresholds $q \in \{0.0, 0.5, 0.9, 0.95, 0.99\}$ as measured by c_{B2}, run on YT-100K with $d = 5$ and $\alpha = 0.05$. GAMINE can ensure $q = 0.5$ with little loss in performance, and it can reduce our objective considerably even under a strict $q = 0.95$.

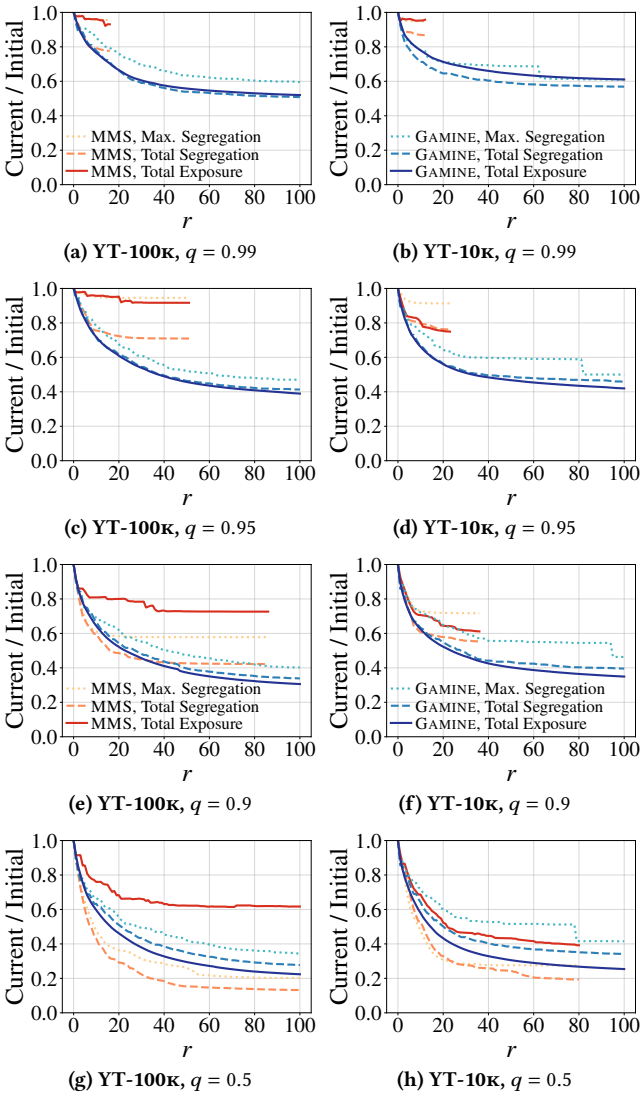


(a) Regular out-degree d = 10 **(b) Regular out-degree d = 5**

Figure 3: Performance of GAMINE with $q = 0.0$, compared with the four baselines BL1, BL2, BL3, and BL4 under c_{B1}, run on YT-100K with $\alpha = 0.05$ and χ = U. As BL4 is roundless, we apply its rewirings in decreasing order of Δ to depict its performance as a function of r. GAMINE outcompetes all baselines, but BL3 and BL4 also show strong performance.

for each of our datasets. This corresponds to the time taken by 1-REM in GAMINE and by 1-REWIRING in MMS, which drives the overall scaling behavior of both algorithms. We find that GAMINE scales approximately linearly, whereas MMS scales approximately quadratically (contrasting with the empirical time complexity of $O(n \log n)$ claimed in [10]). This is because our implementation of MMS follows the original authors', whose evaluation of the segregation objective takes time $O(n)$ and is performed $O(m)$ times. The speed of precomputations depends on the problem variant (REM vs. QREM), and for QREM, also on the quality function θ. In our experiments, precomputations add linear overhead for GAMINE and volatile overhead for MMS, as we report in Appendix F.2.

6.2.4 Data complexity. Given that GAMINE strongly reduces the expected total exposure to harm with few rewirings on the YouTube data, as evidenced in Figs. 2 to 4, one might be surprised to learn that its performance seems much weaker on the NELA-GT data (Appendix F.4): While it still reduces the expected total exposure and outperforms MMS (which struggles to reduce its objective at all on the NF data), the impact of individual rewirings is much smaller than on the YouTube datasets, and the value of the quality threshold q barely makes a difference. This motivates us to investigate how *data complexity* impacts our ability to reduce the expected total exposure to harm via edge rewiring: Could reducing exposure to harm be *intrinsically* harder on NF data than on YT data? The



(a) YT-100K, $q = 0.99$ (b) YT-10K, $q = 0.99$

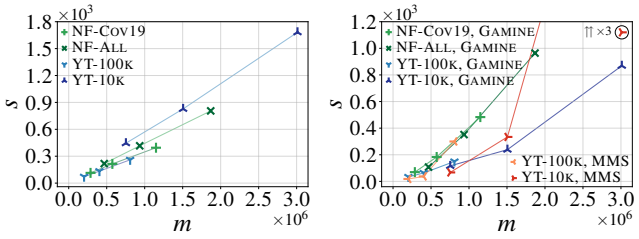
(c) YT-100K, $q = 0.95$ (d) YT-10K, $q = 0.95$

(e) YT-100K, $q = 0.9$ (f) YT-10K, $q = 0.9$

(g) YT-100K, $q = 0.5$ (h) YT-10K, $q = 0.5$

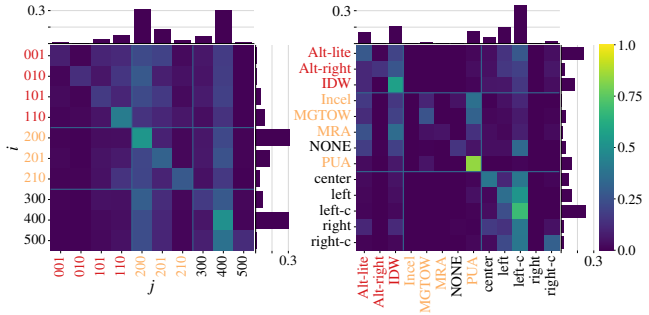
Figure 4: Performance of GAMINE and MMS when measured under c_{B1} by the maximum segregation or the total segregation from Fabbri et al. [10], or by the total exposure as defined in Eq. (3), run on YT-100K (left) and YT-10K (right) with $d = 5$, $\alpha = 0.05$, and $\chi = \mathrm{U}$. For all but $q = 0.5$, GAMINE outperforms MMS on *all* objectives, and MMS stops early because it can no longer reduce the maximum segregation.

answer is yes. First, the in-degree distributions of the YT graphs are an order of magnitude more skewed than those of the NF graphs (Appendix E.2.3, Fig. 15). This is unsurprising given the different origins of their edges (user interactions vs. cosine similarities), but it creates opportunities for high-impact rewirings involving highly prominent nodes in YT graphs (which GAMINE seizes in practice, see below). Second, as depicted in Fig. 6, harmful and benign nodes are much more strongly interwoven in the NF data than in the YT data. This means that harmful content is less siloed in the NF graphs, but it also impedes strong reductions of the expected total exposure.



(a) GAMINE on REM (b) GAMINE and MMS on QREM

Figure 5: Scaling of GAMINE and MMS under c_{B1} with $\alpha = 0.05$, $\chi = \mathrm{U}$, and $q = 0.0$ (REM) resp. 0.99 (QREM). We report the seconds s to compute a single rewiring as a function of $m = dn$ (MMS does not identify any rewirings on NF-COV19 and NF-ALL). GAMINE scales more favorably than MMS.

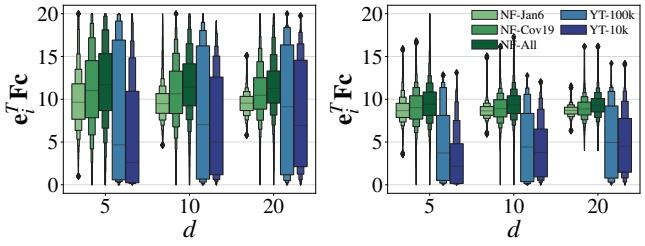


(a) NF-ALL (b) YT-10K

Figure 6: Fractions of edges running between news outlet resp. video channel categories for real-world graphs with $d = 5$, with marginals indicating the fraction of sources (right) resp. targets (top) in each category. News outlet categories are denoted as triples (veracity score, conspiracy-pseudoscience flag, questionable-source flag); for video channel categories, {left, right}-center is abbreviated as {left, right}-c; and label colors are coarse indicators of harm. In NF-ALL, harmful and benign nodes are more interconnected than in YT-10K.

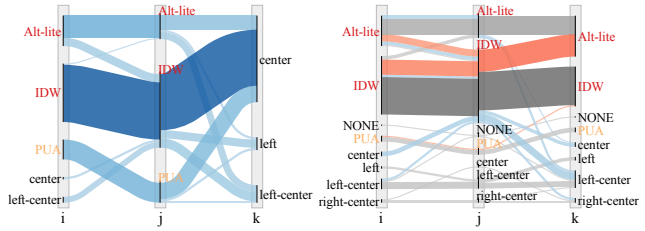
Third, as a result of the two previous properties, the initial node exposures are much more concentrated in the NF graphs than in the YT graphs, as illustrated in Fig. 7, with a median sometimes twice as large as the median of the identically parametrized YT graphs, and a much higher average exposure (cf. $f(G)/n$ in Table 2). Finally, the relevance scores are much more skewed in the YT data than in the NF data (Appendix E.2.3, Fig. 16). Hence, while we are strongly constrained by q on the YT data even when considering only the 100 highest-ranked nodes as potential rewiring targets, we are almost unconstrained in the same setting on the NF data, which explains the comparative irrelevance of q on the NF data. Thus, the performance differences we observe between the NF data and the YT data are due to intrinsic dataset properties: REM and QREM are simply more complex on the news data than on the video data.

6.2.5 General guidelines. Finally, we would like to abstract the findings from our experiments into general guidelines for reducing exposure to harm in recommendation graphs, especially under quality constraints. To this end, we analyze the metadata associated



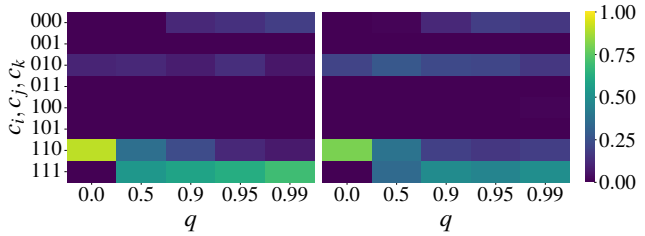
(a) c_{B2}, χ = S (b) c_{R1}, χ = U

Figure 7: Distributions of initial node exposures $\mathbf{e}_i^T\mathbf{Fc}$ in our real-world datasets, computed with $\alpha = 0.05$. Note that cost functions sharing a name are defined differently for the YT and NF datasets (based on their semantics). The NF datasets generally exhibit more concentrated exposure distributions than the YT datasets and higher median exposures.



(a) Quality threshold $q = 0.0$ (b) Quality threshold $q = 0.99$

Figure 8: Channel class of videos in rewirings (i, j, k) on YT-100K with $d = 5$, $\alpha = 0.05$, and χ = U, computed using c_{R1}, for different quality thresholds. Rewirings between classes are color-scaled by their count, using blues if $c_{R1}(k) < c_{R1}(j)$, reds if $c_{R1}(k) > c_{R1}(j)$, and grays otherwise. For $q = 0.0$, we only replace costly targets j by less costly targets k, as expected, but for $q = 0.99$, we see many rewirings with $c_{R1}(k) \geq c_{R1}(j)$.



(a) YT-100K (b) YT-10K

Figure 9: Mapping the nodes in each rewiring (i, j, k) to their costs (c_i, c_j, c_k), we report the fraction of rewirings in each cost class under c_{B1} and $q \in \{0.0, 0.5, 0.9, 0.95, 0.99\}$, for YT graphs with $d = 5$, $\alpha = 0.05$, and χ = U. While most intuitively suboptimal classes occur rarely (e.g., 001, 011, 101), under quality constraints, we often rewire *among* harmful nodes.

with our rewirings. In particular, for each set of rewirings (i, j, k) obtained in our experiments, we are interested in the channel resp. news outlet classes involved, as well as in the distributions of cost triples (c_i, c_j, c_k) and in-degree tuples $(\delta^-(i), \delta^-(j))$. As exemplified in Fig. 8, while we consistently rewire edges from harmful to benign targets in the quality-unconstrained setting ($q = 0.0$), under strict quality control ($q = 0.99$), we frequently see rewirings from harmful to equally or more harmful targets. More generally, as illustrated in Fig. 9, the larger the threshold q, the more we rewire *among* harmful, resp. benign, nodes ($c_i = c_j = c_k = 1$, resp. 0)—which MMS does not even allow. Furthermore, the edges we rewire typically connect nodes with large in-degrees (Appendix F.5, Fig. 28). We conclude that a simplified strategy for reducing exposure to harm under quality constraints is to identify edges that connect high-cost nodes with large in-degrees, and rewire them to the node with the lowest exposure among all nodes meeting the quality constraints.

7 DISCUSSION AND CONCLUSION

We studied the problem of reducing the exposure to harmful content in recommendation graphs by edge rewiring. Modeling this exposure via absorbing random walks, we introduced QREM and REM as formalizations of the problem with and without quality constraints on recommendations. We proved that both problems are NP-hard and NP-hard to approximate to within an additive error, but that under mild assumptions, the greedy method provides a $(1-1/e)$-approximation for the REM problem. Hence, we introduced GAMINE, a greedy algorithm for REM and QREM running in linear time under realistic assumptions on the input, and we confirmed its effectiveness, robustness, and efficiency through extensive experiments on synthetic data as well as on real-world data from video recommendation and news feed applications.

Our work improves over the state of the art (MMS by Fabbri et al. [10]) in terms of performance, and it eliminates several limitations of prior work. While Fabbri et al. [10] model benign nodes as *absorbing* states and consider a brittle *max*-objective that is minimized even by highly harm-exposing recommendation graphs, we model benign nodes as *transient* states and consider a robust *sum*-objective that captures the overall consumption of harmful content by users starting at any node in the graph. Whereas MMS can only handle *binary* node labels, GAMINE works with *real-valued* node attributes, which permits a more nuanced encoding of harmfulness.

We see potential for future work in several directions. For example, it would be interesting to adapt our objective to mitigate *polarization*, i.e., the separation of content with opposing views, with positions modeled as positive and negative node costs. Moreover, we currently assume that all nodes are equally likely as starting points of random walks, which is unrealistic in many applications. Finally, we observe that harm reduction in recommendation graphs has largely been studied in separation from harm reduction in other graphs representing consumption phenomena, such as user interaction graphs. A framework for optimizing functions under budget constraints that includes edge rewirings, insertions, and deletions could unify these research lines and facilitate future progress.

ACKNOWLEDGMENTS

This research is supported by the Academy of Finland project MLDB (325117), the ERC Advanced Grant REBOUND (834862), the EC H2020 RIA project SoBigData++ (871042), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] Rediet Abebe, T-H Hubert Chan, Jon Kleinberg, Zhibin Liang, David Parkes, Mauro Sozio, and Charalampos E Tsourakakis. 2021. Opinion dynamics optimization by varying susceptibility to persuasion via non-convex local search. *ACM Transactions on Knowledge Discovery from Data* 16, 2 (2021), 1–34.

[2] Florian Adriaens, Honglian Wang, and Aristides Gionis. 2023. Minimizing hitting time between disparate groups with shortcut edges. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. To appear.

[3] Victor Amelkin and Ambuj K Singh. 2019. Fighting opinion control in social networks via link recommendation. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. 677–685.

[4] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. In *Proceedings of the ACM on Human-Computer Interaction (CHI)*, Vol. 5. 1–34.

[5] Hau Chan and Leman Akoglu. 2016. Optimizing network robustness by edge rewiring: a general framework. *Data Mining and Knowledge Discovery* 30, 5 (2016), 1395–1425.

[6] Uthsav Chitra and Christopher Musco. 2020. Analyzing the impact of filter bubbles on social network polarization. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 115–123.

[7] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[8] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. 2014. Modeling opinion dynamics in social networks. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 403–412.

[9] Paul Erdős and Alfréd Rényi. 1959. On random graphs I. *Publicationes Mathematicae* 6, 1 (1959), 290–297.

[10] Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. 2022. Rewiring what-to-watch-next recommendations to reduce radicalization pathways. In *Proceedings of the ACM Web Conference*. 2719–2728.

[11] Uriel Feige. 2003. Vertex cover is hardest to approximate on regular graphs. Technical Report MCS03–15.

[12] Antonio Ferrara, Lisette Espín-Noboa, Fariba Karimi, and Claudia Wagner. 2022. Link recommendations: Their impact on network structure and minorities. In *Proceedings of the ACM Web Science Conference*. 228–238.

[13] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2017. Reducing controversy by connecting opposing views. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 81–90.

[14] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[15] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. 2013. Opinion maximization in social networks. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*. 387–395.

[16] Raymond Greenlaw and Rossella Petreschi. 1995. Cubic graphs. *ACM Computing Surveys (CSUR)* 27, 4 (1995), 471–495.

[17] Mauricio Gruppi, Benjamin D. Horne, and Sibel Adali. 2021. NELA-GT-2021: A large multi-labelled news dataset for the study of misinformation in news articles. arXiv:2203.05659 [cs.CY]

[18] Shahrzad Haddadan, Cristina Menghini, Matteo Riondato, and Eli Upfal. 2022. Reducing polarization and increasing diverse navigability in graphs by inserting edges and swapping edge weights. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2334–2378.

[19] Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences* 118, 32 (2021), e2101967118.

[20] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. 263–272.

[21] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on Human-Computer Interaction (CHI)* 4, CSCW1 (2020), 1–27.

[22] Victor P Il'ev. 2001. An approximation guarantee of the greedy descent algorithm for minimizing a supermodular set function. *Discrete Applied Mathematics* 114, 1-3 (2001), 131–146.

[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[24] Subhash Khot. 2002. On the power of unique 2-prover 1-round games. In *Proceedings of the Annual ACM Symposium on Theory of Computing*. 767–775.

[25] Subhash Khot and Oded Regev. 2008. Vertex cover might be hard to approximate to within $2-\varepsilon$. *J. Comput. System Sci.* 74, 3 (2008), 335–349.

[26] Baris Kirdemir and Nitin Agarwal. 2022. Exploring bias and information bubbles in YouTube's video recommendation networks. In *Proceedings of the International Conference on Complex Networks and Their Applications*. 166–177.

[27] Mark Ledwich and Anna Zaitsev. 2020. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday* (2020).

[28] Mark Ledwich, Anna Zaitsev, and Anton Laukemper. 2022. Radical bubbles on YouTube? Revisiting algorithmic extremism with personalised recommendations. *First Monday* (2022).

[29] Robin Mamié, Manoel Horta Ribeiro, and Robert West. 2021. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In *Proceedings of the ACM Web Science Conference*. 139–147.

[30] Charalampos Mavroforakis, Michael Mathioudakis, and Aristides Gionis. 2015. Absorbing random-walk centrality: Theory and algorithms. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. 901–906.

[31] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (2001), 415–444.

[32] Sourav Medya, Arlei Silva, Ambuj Singh, Prithwish Basu, and Ananthram Swami. 2018. Group centrality maximization via network design. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*. 126–134.

[33] Marco Minici, Federico Cinus, Corrado Monti, Francesco Bonchi, and Giuseppe Manco. 2022. Cascade-based echo chamber detection. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 1511–1520.

[34] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming* 14, 1 (1978), 265–294.

[35] Lutz Oettershagen, Petra Mutzel, and Nils M Kriege. 2022. Temporal walk centrality: Ranking nodes in evolving networks. In *Proceedings of the ACM Web Conference*. 1640–1650.

[36] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2022. "It is just a flu": Assessing the effect of watch history on YouTube's pseudoscientific video recommendations. In *Proceedings of the International AAAI Conference on Web and Social Media*. 723–734.

[37] Nikos Parotsidis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2015. Selecting shortcuts for a smaller world. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*. 28–36.

[38] Bibek Paudel and Abraham Bernstein. 2021. Random walks with erasure: Diversifying personalized recommendations on social and information networks. In *Proceedings of the ACM Web Conference*. 2046–2057.

[39] Niccolo Pescetelli, Daniel Barkoczi, and Manuel Cebrian. 2022. Bots influence opinion dynamics without direct human-bot interaction: The mediating role of recommender systems. *Applied Network Science* 7, 1 (2022), 1–19.

[40] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 3982–3992.

[41] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 131–141.

[42] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. In *Proceedings of the ACM on Human-Computer Interaction (CHI)*, Vol. 2. 1–22.

[43] Jack Sherman and Winifred J Morrison. 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* 21, 1 (1950), 124–127.

[44] Larissa Spinelli and Mark Crovella. 2017. Closed-loop opinion formation. In *Proceedings of the ACM Web Science Conference*. 73–82.

[45] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, et al. 2022. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems* (2022).

[46] Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Konstantinos Semertzidis, and Panayiotis Tsaparas. 2022. Link recommendations for PageRank fairness. In *Proceedings of the ACM Web Conference*. 3541–3551.

[47] Antoine Vendeville, Anastasios Giovanidis, Effrosyni Papanastasiou, and Benjamin Guedj. 2023. Opening up echo chambers via optimal content recommendation. In *Proceedings of the International Conference on Complex Networks and Their Applications*. 74–85.

[48] Tomasz Wąs, Talal Rahwan, and Oskar Skibski. 2019. Random walk decay centrality. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2197–2204.

[49] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review* 10, 2 (2021), 1–29.

[50] Muhsin Yesilada and Stephan Lewandowsky. 2022. Systematic review: YouTube recommendations and problematic content. *Internet Policy Review* 11, 1 (2022), 1–22.

[51] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. 2010. A unified framework for link recommendation using random walks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 152–159.
[52] Xiaojuan Zhang, Qian Liu, Min Li, and Yang Zhou. 2022. Fast algorithms for supermodular and non-supermodular minimization via bi-criteria strategy. *Journal of Combinatorial Optimization* 44, 5 (2022), 3549–3574.
[53] Liwang Zhu, Qi Bao, and Zhongzhi Zhang. 2021. Minimizing polarization and disagreement in social networks via link recommendation. *Advances in Neural Information Processing Systems*, 2072–2084.
[54] Liwang Zhu and Zhongzhi Zhang. 2022. A nearly-linear time algorithm for minimizing risk of conflict in social networks. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (KDD)*. 2648–2656.

ETHICS STATEMENT

In this work, we introduce Gamine, a method to reduce the exposure to harm induced by recommendation algorithms on digital media platforms via edge rewiring, i.e., replacing certain recommendations by others. While removing harm-inducing recommendations constitutes a milder intervention than censoring content directly, it still steers attention away from certain content to other content, which, if pushed to the extreme, can have censorship-like effects. Although in its intended usage, Gamine primarily counteracts the tendency of recommendation algorithms to overexpose harmful content as similar to other harmful content, when fed with a contrived cost function, it could also be used to discriminate against content considered undesirable for problematic reasons (e.g., due to political biases or stereotypes against minorities). However, as the changes to recommendations suggested by Gamine could also be made by amending recommendation algorithms directly, the risk of *intentional* abuse is no greater than that inherent in the recommendation algorithms themselves, and *unintentional* abuse can be prevented by rigorous impact assessments and cost function audits before and during deployment. Thus, we are confident that overall, Gamine can contribute to the health of digital platforms.

APPENDIX

In addition to Table 3, included below, the written appendix to this work contains the following sections:

A Omitted proofs
B Other graph edits
C Omitted pseudocode
D Reproducibility information
E Dataset information
F Further experiments

This appendix, along with the main paper, is available on arXiv and also deposited at the following DOI: 10.5281/zenodo.8002980. To facilitate reproducibility, all code, data, and results are made available at the following DOI: 10.5281/zenodo.7936816.

Table 3: Most important notation used in this work.

Symbol	Definition	Description
		Graph Notation
G	$= (V, E)$	Graph
n	$= \lvert V \rvert$	Number of nodes
m	$= \lvert E \rvert$	Number of edges
$\delta^-(i)$	$= \lvert \{j \mid (j,i) \in E\} \rvert$	In-degree of node i
$\Gamma^+(i)$	$= \{j \mid (i,j) \in E\}$	Set of out-neighbors of node i
$\delta^+(i)$	$= \lvert \Gamma^+(i) \rvert$	Out-degree of node i
d		Regular out-degree of an out-regular graph
Δ^+	$= \max\{\delta^+(i) \mid i \in V\}$	Maximum out-degree
S	$= \{i \in V \mid \mathbf{e}_i^T \mathbf{F} \mathbf{c} = 0\}$	Set of safe nodes
U	$= \{i \in V \mid \mathbf{e}_i^T \mathbf{F} \mathbf{c} > 0\}$	Set of unsafe nodes
Λ^+	$= \max\{\delta^+(i) \mid i \in U\}$	Maximum out-degree of an unsafe node
		Matrix Notation
$\mathbf{M}[i,j]$		Element in row i, column j of $\mathbf{M}$
$\mathbf{M}[i,:]$		Row i of $\mathbf{M}$
$\mathbf{M}[:,j]$		Column j of $\mathbf{M}$
$\mathbf{e}_i$		i-th unit vector
$\mathbf{1}$		All-ones vector
$\mathbf{I}$		Identity matrix
$\lVert \mathbf{M} \rVert_\infty$	$= \max_i \sum_{j=0}^{n} \mathbf{M}[i,j]$	Infinity norm
		Notation for REM and QREM
(i,j,k)		Rewiring replacing $(i,j) \in E$ by $(i,k) \notin E$ with $p_{ik} = p_{ij}$, cf. Table 1
r	$\in \mathbb{N}$	Rewiring budget
α	$\in (0,1]$	Random-walk absorption probability
p_{ij}	$\in (0, 1-\alpha]$	Probability of traversing (i,j) from i
$\mathbf{P}$	$\in [0, 1-\alpha]^{n \times n}$	Random-walk transition matrix
$\mathbf{F}$	$= \sum_{i=0}^{\infty} \mathbf{P}^i = (\mathbf{I} - \mathbf{P})^{-1}$	Fundamental matrix
c		Cost function with range $[0,1]$
c_i	$\in [0,1]$	Cost associated with node i
$\mathbf{c}$	$\in [0,1]^n$	Vector of node costs
κ	$\in \mathbb{N}$	Number of power iterations
χ	$\in \{\mathbf{U}, \mathbf{S}\}$	Shape of probability distribution over the out-edges of a node
		Notation for QREM Only
$\mathbf{R}$	$\in \mathbb{R}_{\geq 0}^{n \times n}$	Relevance matrix
θ		Relevance function with range $[0,1]$
q	$\in [0,1]$	Quality threshold
$\mathbf{r}_i$	$\in V^{\delta^+(i)}$	Relevance-ordered targets of out-edges of i
$\mathrm{idx}_i(j)$		Relevance rank of node j for node i
$T_{\delta^+}(i)$	$= \{j \mid \mathrm{idx}_i(j) \leq \delta^+(i)\}$	Set of the $\delta^+(i)$ nodes most relevant for node i
DCG	$= \sum_{j \in \Gamma^+(i)} \frac{\mathbf{R}[i,j]}{\log_2(1+\mathrm{idx}_i(j))}$	Discounted Cumulative Gain
iDCG	$= \sum_{j \in T_{\delta^+}(i)} \frac{\mathbf{R}[i,j]}{\log_2(1+\mathrm{idx}_i(j))}$	Ideal Discounted Cumulative Gain
nDCG	$= \frac{\mathrm{DCG}(i)}{\mathrm{iDCG}(i)}$	Normalized Discounted Cumulative Gain
		Notation Related to the Exposure Function f and its Analysis
$f(G)$	$= \mathbf{1}^T \mathbf{F} \mathbf{c}$	Exposure function (minimization objective)
$f_\Delta(G, G_r)$	$= f(G) - f(G_r)$	Reduction-in-exposure function (equivalent maximization objective)
$G', \mathbf{P}', \mathbf{F}'$		Graph G, transition matrix $\mathbf{P}$, fundamental matrix $\mathbf{F}$, as updated by rewiring (i,j,k), cf. Table 1
$\mathbf{u}$	$= p_{ij} \mathbf{e}_i$	Vector capturing the source i of a rewiring (i,j,k) and the traversal probability of (i,j)
$\mathbf{v}$	$= \mathbf{e}_j - \mathbf{e}_k$	Vector capturing the old target j and the new target k of a rewiring (i,j,k)
σ	$= \mathbf{1}^T \mathbf{F} \mathbf{u}$	p_{ij}-scaled i-th column sum
τ	$= \mathbf{v}^T \mathbf{F} \mathbf{c}$	$\mathbf{c}$-scaled sum of differences between the j-th row sum and the k-th row sum
ρ	$= 1 + \mathbf{v}^T \mathbf{F} \mathbf{u}$	Normalization factor ensuring that $\mathbf{F}'\mathbf{1} = \mathbf{F}\mathbf{1}$
Δ	$= f_\Delta(G, G') = \sigma\tau/\rho$	Reduction of f obtained by a single rewiring (i,j,k)
$\hat{\Delta}$	$= \Delta\rho = \sigma\tau$	Heuristic for Δ