Computationally Establishing Causality: The Impact of Demographic Minorities on Bias in Opinion Diffusion Models

Michalina Janik[1*], **Radosław Michalski**[2], **Akhil Arora**[1],

[1]Department of Computer Science, Aarhus University, Denmark
[2]Department of Artificial Intelligence, Wrocław University of Science and Technology, Poland
au749513@uni.au.dk, akhil.arora@cs.au.dk, radoslaw.michalski@pwr.edu.pl

Abstract

Network-based opinion diffusion models often assume uniform influence mechanisms in diverse subpopulations. This assumption risks embedding biases against demographic minorities whose interaction patterns diverge from those of majority groups. To rigorously assess whether minorities causally affect model accuracy, we propose a computationally grounded experimental approach. Using the CoDiNG opinion diffusion model and real-world interaction data from the NetSense study, we construct counterfactual networks by systematically removing minority nodes – identified by attributes such as ethnicity, socioeconomic status, and gender – while preserving network structure through degree-preserving rewiring as well as temporal integrity of events stream. Empirical results show that removing minorities leads to a 29% drop in F1-score (from 0.45 ± 0.11 to 0.32 ± 0.07), loss of polarization sensitivity, and significantly flatter response curves. Minority nodes also exhibit disproportionately higher centrality (e.g., degree: 9.01 vs 6.50 for income groups), confirming their structural importance. Our findings provide causal evidence that demographic minorities are not peripheral outliers but critical to maintaining the fidelity of opinion diffusion models.

Introduction

Opinion diffusion models are widely used in computational social science to understand how beliefs propagate through social networks. However, these models often assume uniform influence mechanisms across all nodes, overlooking the distinct structural and behavioral roles played by minority subgroups. This oversight risks embedding systematic biases that degrade model fidelity, particularly when modeling diverse populations.

Prior work emphasizes the role of network topology in shaping diffusion outcomes (McPherson, Smith-Lovin, and Cook 2001; Centola 2010; Aral and Walker 2012), yet the causal contribution of minority groups—especially those defined by demographic status—remains underexplored. Although theories of homophily and minority influence suggest these groups may play disproportionate roles in bridging or reinforcing subnetwork opinions, existing models rarely evaluate their contributions using rigorous counterfactual or causal frameworks.

To address this gap, we develop a computational methodology for isolating the influence of minority groups in networked opinion diffusion. Using temporally resolved interaction data from the NetSense study (Striegel et al. 2013) and the cognitively grounded CoDiNG model (Nurek et al. 2024), we systematically remove minority nodes—defined by attributes such as parental income, religion, or education—and reconstruct counterfactual networks using degree-preserving rewiring and validated temporal alignment. This design enables robust evaluation of causal effects on opinion spread and model performance.

To ensure conceptual precision, we distinguish between two types of minority status considered in this study. *Demographic minorities* refer to individuals underrepresented in terms of structural attributes such as parental income, education level, or religion. In contrast, we define *behavioral minorities* as numerically rare groups based on preference or attitudinal features—for example, users with restrictive Facebook privacy settings. While both categories involve underrepresentation, only demographic minorities are hypothesized to exert structural influence on opinion dynamics due to their embeddedness in broader patterns of social inequality. This distinction allows us to evaluate whether observed model degradation and centrality shifts arise from structural marginality or merely from low numerical prevalence.

We formally evaluate the following hypotheses:

- **H1 (Structural Centrality Hypothesis):** Demographic minority nodes exhibit significantly higher centrality (e.g., degree, betweenness, or closeness) compared to demographic majority nodes within the same network.
- **H2 (Behavioral Adaptability Hypothesis):** Demographic minority nodes display greater longitudinal opinion variability (i.e., higher opinion fluidity) than demographic majority nodes across survey waves.
- **H3 (Structural Bridging Hypothesis):** Demographic minority nodes form significantly more cross-group (demographic boundary-spanning) edges than demographic majority nodes, acting as structural bridges.
- **H4 (Causal Performance Hypothesis):** Removing demographic minority nodes from the network causes a sig-

*Corresponding author: Michalina Janik, au749513@uni.au.dk

nificant drop in opinion diffusion model performance (F1 score), even when degree distribution and temporal dynamics are preserved.

Related Work

Opinion Diffusion Models

Opinion diffusion models are a convenient way to model how opinions spread throughout society. Initially, before the rise of network-oriented approaches,this was typically modeled using macro-level diffusion equations or spatial autocorrelation techniques with notable examples of parametric logistic model (Carrington, Scott, and Wasserman 2005), Bass model (Bass 1969) or Moran's I (Moran 1950). However, nowadays the spread of opinions is mapped on top of a social network which has been built using data coming from surveying individuals or parsing data on communication.

In the network area, the opinion models followed two directions: discrete models and continuous models of opinions. For the former, it was assumed that the opinion can belong to a fixed set of choices, while in the continuous models the agent could be located somewhere in between. Discrete opinion models include the Ising model (Ising 1925), Sznajd model (Sznajd-Weron and Sznajd 2000), threshold models (Valente 1996), and the independent cascade model (Goldenberg, Libai, and Muller 2001). In contrast, continuous opinion models cover the De-Groot model (DeGroot 1974), Deffuant-Weisbuch model (Deffuant et al. 2000) and the Hegselmann-Krause model (Rainer and Krause 2002), among others. However, as researchers tried to incorporate the cognitive processes in how we internally process information with how we externalize opinions, hybrid models combining discrete and continuous approaches are also explored, such as CODA (Martins 2008), SNOAE (Zhan et al. 2021), or SJBO (Fan and Pedrycz 2016). However, these models did not account for mixed opinions, so the CoDiNG model addressed that (Nurek et al. 2024). It is used in this study and is described in more detail in Section Methodology.

Fairness in Social Networks

While machine learning fairness has received increasing attention across classification and prediction tasks, fairness in network-based learning remains relatively underexplored. To provide the best results in tasks such as classification or forecasting, the questions arose if these algorithms are fair with respect to many criteria, starting with biases related to ethnicity, gender, social class, and others. However, as the authors of (Saxena, Fletcher, and Pechenizkiy 2024) note, one of the areas in which much research was not done on algorithmic fairness was on social networks. Here, the algorithms proposed for network clustering, social influence maximization, or link prediction are still in the infancy phase and require more work compared to typical machine learning approaches (Karimi, Oliveira, and Strohmaier 2022).

Previous studies emphasize the role of network structure in shaping opinion diffusion (Centola 2010; McPherson, Smith-Lovin, and Cook 2001). Theories of homophily and minority influence suggest that underrepresented groups exhibit unique diffusion behaviors, potentially affecting model performance. Although some fairness-oriented algorithms attempt to mitigate these biases, empirical evaluations remain limited (Hofstra et al. 2017). Our work extends this research by introducing a computational framework to causally examine minority effects in opinion diffusion models. Unlike prior work that addresses fairness via post hoc mitigation or predictive debiasing, our approach introduces a counterfactual, network-aware causal framework that evaluates the structural necessity of minority presence for fidelity in cognitive opinion models.

Methodology

The CoDiNG Opinion Diffusion Model

The CoDiNG model (Continuous-Discrete Naming Game) (Nurek et al. 2024) integrates cognitive principles into the diffusion of opinions. Contrary to discrete models, it is a combined opinion model that combines two layers: exposed discrete opinions being in line with the Naming Game model (Lu, Korniss, and Szymanski 2009) and a latent layer that is modeled as a continuous sentiment towards binary opinions. This model is more aligned with cognitive processes (Michalski et al. 2021) as the typical mechanics of Naming Game do not account for more subtle changes of opinions that are not externalized. The mechanics of the model tracks the difference between opinions A and B in the latent layer. If this difference is small (less than Δ_o), an agent exposes a mixed opinion AB, but if it is larger than Δ_o, the agent expresses either opinion A or opinion B. In (Nurek et al. 2024) the authors evaluated the model with respect to its precision in modeling opinions among students in the Netsense study and found that the model is capable of modeling the cognitive layer with good results. The mechanics of the model is presented in Figure 1.

The NetSense Study

For the experimental part, we utilize the NetSense dataset (Striegel et al. 2013) – the result of a longitudinal study of University of Notre Dame students. The dataset includes:

- **Structural data**: This component consists of detailed records of communication events, calls and texts, capturing both the timing (timestamps) and the participants (sender and receiver IDs).
- **Survey responses**: These are biannual surveys where students expressed their views on a range of societal and political issues. For our analysis, we specifically focus on six questions that have shown notable fluctuations in opinions across different semesters. These questions are:
  1. **Euthanasia (euthanasia)**: Should physicians be allowed to end a patient's life by painless means upon request when suffering from an incurable disease?
  2. **Social security spending (fssocsec)**: Should federal spending on social security be increased, maintained, or decreased?
  3. **Welfare spending (fswelfare)**: Should federal welfare spending be increased, maintained, or decreased?

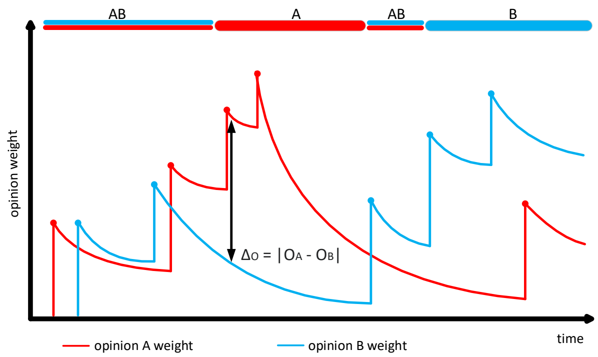

Figure 1: An example of changing opinion in the CoDiNG model (Nurek et al. 2024). If the difference in latent continuous opinion layer is smaller than Δ_o, an agent exposes a discrete mixed opinion AB. Otherwise, it is either opinion A or opinion B being communicated. The discrete opinion that is communicated based on the continuous values of opinions is presented in the top part of the figure.

4. **Government job guarantees (jobguar)**: To what extent should the government ensure that every individual has a job and a good standard of living versus allowing individuals to succeed on their own?
5. **Legality of marijuana (marijuana)**: Should the use of marijuana be legalized?
6. **Equal rights (toomucheqrights)**: Has the push for equal rights in the country gone too far?

These selected issues range from personal ethics to broader social policy debates, making them ideal for analyzing opinion dynamics. The responses were initially recorded on various scales, including binary (yes/no) and multiple levels of agreement. To enable compatibility with the ternary output of the CoDiNG model (agree/disagree/not sure), we standardized survey responses as follows:

– For 6-point Likert scales (e.g., equal rights), responses 5–6 were coded as "agree," 1–2 as "disagree," and 3–4 as "not sure."

– For Yes/No/Don't Know questions (e.g., marijuana, euthanasia), responses were mapped directly to agree/disagree/not sure.

– For policy direction questions (e.g., social security, welfare), "increase" = agree, "decrease" = disagree, and "stay the same" or "don't know" = not sure.

Netsense dataset provides a rich foundation for studying real-world opinion dynamics, as it covers a time span of 6 semesters between years 2011 - 2013, nearly 200 students surveyed multiple times, and more than seven million phone interaction metadata over the course of two and a half years.

Minority Definitions and Selection Criteria

To evaluate the structural and behavioral contributions of underrepresented groups to opinion diffusion, we identify two analytically distinct forms of minority status: *demographic minorities* and *behavioral minorities*. Minority status is defined *within* independently of each variable, rather than based on the total share of the population. This design enables disaggregated analysis and avoids confounding effects due to unequal attribute distributions. For example, a student can be classified as a minority in terms of parental income but not in terms of gender or religion. This attribute-specific framing supports structural precision and avoids overgeneralizing "minority" as a monolithic identity.

Demographic minorities are defined as individuals whose attributes reflect structural dimensions of social inequality and underrepresentation. These include:

- **Parental income (pincome):** The self-reported income bracket of the student's household.
- **Parental education level:**
  - *Mother's education (momed)*
  - *Father's education (daded)*
- **Parental religious affiliation:**
  - *Mother's religion (momrelig)*
  - *Father's religion (dadrelig)*

Behavioral minorities, by contrast, are defined based on attitudinal or preference-based attributes rather than structural inequality. As a behavioral comparator group, we include:

- **Facebook privacy setting (fbprivacy):** Users who report more restrictive privacy preferences than the population median.

For all attributes, a group is classified as a *minority* if it constitutes less than 30% of the population. This consistent threshold enables comparative analysis across diverse attributes while ensuring sufficient statistical power in counterfactual simulations.

This distinction allows us to test whether the structural and functional importance of minorities in opinion diffusion is driven by broader social marginality (demographic minorities) or simply by numerical rarity (behavioral minorities). The inclusion of a non-demographic comparator (Facebook privacy) strengthens the internal validity of our causal claims by offering a contrast group not embedded in systemic social hierarchies.

Minority Group Independence. To verify that our findings are not driven by overlapping individuals across minority definitions, we computed pairwise Jaccard indices between the node sets of each group showed in Figure 2. The average overlap is low (mean Jaccard = 0.09), with the highest observed similarity (0.22) between parental education and religion. This moderate overlap likely reflects underlying sociocultural correlations—such as the co-occurrence of lower educational attainment and minority religious affiliation within families. However, the overlap remains well below levels that would indicate redundant node sets, and each

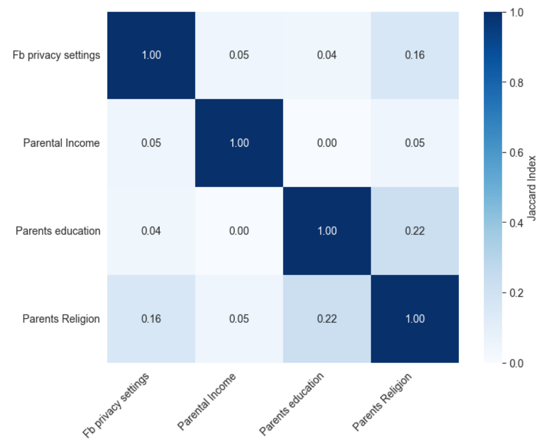

Figure 2: Heatmap of pairwise Jaccard indices between the minority groups.

minority group is removed independently in counterfactual simulations. Other pairs show minimal to no overlap, confirming that observed effects stem from structurally distinct subpopulations.

Counterfactual Network Construction

To isolate the causal impact of the presence of minority on the performance of the model, we construct counterfactual networks by systematically removing nodes identified as demographic minorities. After removal of the nodes, we apply a carefully designed degree-preserving rewiring algorithm (Maslov and Sneppen 2002) to reassign the edges originally attached to the removed nodes, ensuring minimal distortion of the topological characteristics of the network. The preservation of the network topology is critical, as structural characteristics significantly influence diffusion dynamics and can confound causal interpretations (Aral and Walker 2012; Centola 2010; Holme and Saramäki 2012). Specifically, we ensure that the following network metrics are rigorously preserved and statistically validated before performing opinion diffusion simulations.

1. **Degree distribution**: the degree distribution captures node connectivity patterns that strongly influence diffusion dynamics. Degree-preserving rewiring ensures that nodes retain their original degree, maintaining the balance between highly connected and peripheral nodes (Maslov and Sneppen 2002; Newman 2003).
2. **Clustering coefficient**: The clustering coefficient quantifies the density of the local network and the propensity of the network nodes to form tightly knit communities. It has been shown to significantly affect the adoption rate and overall effectiveness of complex behavioral diffusion (Centola 2010; Watts and Strogatz 1998). Centola's experimental results explicitly show that higher clustering leads to greater behavioral adoption due to social reinforcement, highlighting the necessity of preserving clustering when modifying the topology (Centola 2010).
3. **Network modularity**: Modularity measures the strength of the division of the network into modules or communities, capturing the community structure that affects the diffusion pathways (Newman and Girvan 2004). Preserving modularity ensures that the rewiring process does not artificially create or disrupt communities, which could influence the dynamics of opinion propagation in unintended ways.

After rewiring, we performed Kolmogorov-Smirnov (KS) tests on the degree distribution and the clustering coefficient to confirm that the resulting network retains its original structural integrity. Only counterfactual networks that meet these criteria are utilized for further simulations to ensure robust and causally valid interpretations.

This approach strikes a balance between ensuring topological validity while avoiding infeasible constraints on diameter and path length, which are highly sensitive to any node removal (Holme and Saramäki 2012).

Modification of the Event Stream

When rewiring networks to create counterfactual scenarios involving the removal of demographic minorities, it is critical not only to preserve structural properties (degree and clustering) but also to maintain temporal realism. Temporal integrity in event streams, reflecting realistic patterns of interactions over time, is crucial for accurate simulations of opinion diffusion processes, particularly using temporal-sensitive models such as CoDiNG (Holme and Saramäki 2012; Aral and Walker 2012).

Previous research emphasizes that temporal network properties, such as event bursts and time correlations, significantly affect diffusion dynamics, opinion formation, and network stability (Holme and Saramäki 2012; Masuda and Holme 2020). In particular, Holme and Saramäki (Holme and Saramäki 2012) highlight that disregarding temporal structure can lead to substantial misrepresentations of diffusion processes, undermining the validity of causal inferences derived from computational experiments.

Therefore, to modify the event stream without introducing temporal distortions, we follow an explicit two-step procedure. First, we remove events involving deleted minority nodes. Second, we employ explicit edge mapping techniques to substitute these removed events with temporally equivalent interactions among the remaining nodes. This approach ensures that the modified event stream preserves the original temporal distributions and structural coherence of interactions, a practice consistent with the state-of-the-art temporal network research (Aral and Walker 2012; Centola 2010; Maslov and Sneppen 2002).

Finally, we quantitatively validate the temporal preservation using Kolmogorov–Smirnov statistical tests on event timing distributions before and after modification (Aral and Walker 2012). Only event streams that pass these rigorous statistical validations are subsequently used for opinion diffusion simulations.

Evaluation and Experimental Design

This section details how we assess the impact of the removal of demographic minority on the diffusion of opinions within simulated networks. Our approach is designed to balance the rigor of experimental controls with the practical realities of network behavior under manipulation.

Experimental Setup

Our experiments are structured to methodically alter the demographic composition of the network by removing nodes associated with specific minority groups and subsequently analysing the effects on the diffusion of opinions. Each network in the NetSense dataset is adjusted per minority group and tested in six different opinion questions described in Section Methodology, thus creating multiple unique experimental conditions, specifically generating one event stream per adjusted graph without a given minority described and specified in the same section.

Network Manipulation and Metrics Preservation

To isolate the causal impact of the presence of minorities on the diffusion of opinions, we first remove nodes identified as belonging to specific demographic minorities. After removal, we reconstruct the network by strategically rewiring edges to preserve its key topological properties. Unlike traditional random rewiring methods, our approach approximates the original connectivity patterns of the removed nodes to maintain structural fidelity.

We adopt empirically grounded tolerance thresholds to evaluate the validity of each counterfactual network:

- **Degree preservation**: We enforce that node degrees remain within a ±5% margin of their original values. This constraint preserves the influence potential of individual nodes and ensures that rewired networks retain comparable connectivity profiles.
- **Clustering coefficient and modularity**: We allow up to 15% deviation in clustering coefficient and modularity relative to the original network. This level of tolerance balances the need to preserve meso-scale structures (e.g., community density and modular boundaries) with the feasibility of generating rewired networks after minority node removal.
- **Statistical validation**: We further apply Kolmogorov-Smirnov tests to confirm that the rewired network's degree distribution and clustering remain statistically comparable to the original. Only networks that pass these tests and fall within the above tolerances are used in downstream simulations.

These thresholds were selected through iterative testing to ensure structural comparability while minimizing the rejection rate of rewired candidates. Stricter bounds (e.g., ±2%) proved infeasible, while looser constraints (e.g., ±20%) introduced excessive topological drift that compromised causal interpretability.

Event Stream Generation for CoDiNG Simulations

After constructing counterfactual networks through minority node removal and degree-preserving rewiring, we generate modified event streams that align with the new network topology while preserving temporal patterns. This process

Algorithm 1: Event Stream Adjustment with Temporal Validation

Require: Original event stream E, Edge mapping M from rewiring
Ensure: Adjusted event stream E', Temporal integrity flag
$E' \leftarrow \text{copy}(E)$
for each event $(u, v, t) \in E'$ **do**
  if $(u, v) \in M$ **then**
    $(u', v') \leftarrow M[(u, v)]$
    Update event to (u', v', t)
  else if $(v, u) \in M$ **then**
    $(v', u') \leftarrow M[(v, u)]$
    Update event to (u', v', t)
  end if
end for
$D_{\text{orig}} \leftarrow \text{KS-test}(E_{\text{time}}, E'_{\text{time}})$
if $p_{\text{value}} > 0.05$ **then**
  return E', True
else
  return E', False
end if

ensures that the CoDiNG model receives realistic interaction data that reflect both the structural and temporal properties of the original network.
The adjustment process follows three key phases:

1. **Edge Mapping Application**:
   - Replace sender/receiver pairs in original events using the edge mapping generated during network rewiring
   - Preserve original timestamps while updating participant IDs
   - Handle both directional ((u, v) and (v, u)) edge relationships
2. **Temporal Pattern Preservation**:
   - Maintain original event timing distributions through non-destructive timestamp preservation
   - Validate using Kolmogorov-Smirnov test on inter-event times:
   $$D_{\text{KS}} = \sup_t |F_{\text{orig}}(t) - F_{\text{adj}}(t)| \quad (1)$$
   - Require $p_{\text{value}} > 0.05$ for temporal integrity
3. **CoDiNG Input Generation**:
   - Export validated event streams in temporal interaction format
   - Maintain original event ordering and burstiness patterns
   - Append demographic metadata for opinion initialization

This approach ensures that modified event streams:

- Maintain temporal correlations critical for opinion cascade formation
- Preserve interaction frequency distributions
- Align with the rewired network's connectivity patterns

Simulation of Opinion Dynamics

Utilizing the CoDiNG model, we simulate the dynamics of opinion under each network condition to observe how changes in the network structure affect the spread of opinion. The model parameters are set as follows:

- **Initial opinion weights**: Result of survey responses, with weights assigned based on agreement, disagreement, or neutrality toward specific issues.
- **Model parameters**: $\mu = 0.3$, $\theta = 0.2$, and $\lambda = 0.005631$, as specified in the original CoDiNG study, to ensure consistency with established cognitive dynamics. For Δ, we evaluated 0.1 to 0.9 with a step of 0.1.

Metric Evaluation

The primary metric for evaluating the impact of network modifications on opinion diffusion is the F1 score, which balances the precision and recall of the model's predictions. This metric is particularly suitable for comparing our results with those from baseline conditions (unmodified network states) and the original CoDiNG study findings.

Statistical Analysis and Result Validation

To ascertain the significance of our findings, pairwise statistical tests are performed on the average F1 score achieved in 10 iterations for the same setting of a delta parameter of the CoDiNG model. The tested delta values range from 0.1 to the 0.9 value, gradually increasing by 0.1.

- For each pair of results associated with the same lambda parameter, we first perform a Shapiro-Wilk test on the differences between the original and modified network outputs to check for normality. This step determines whether the data distributions support the use of parametric methods.
- Depending on the outcome of the normality test, we employ the appropriate statistical test:
  - if data is normally distributed, a paired t-test is conducted to detect statistically significant differences between the outputs from original and modified networks. This test helps validate the causal effects of demographic alterations;
  - if data does not show normal distribution, a Wilcoxon signed-rank test is used as a non-parametric alternative to evaluate the same differences.
- Ten iterations of each simulation scenario are run to address the inherent randomness of the model. The results of these iterations are averaged to ensure the robustness and reliability of the findings.

Results

Overview. This section presents empirical findings corresponding to the four hypotheses introduced in Section . We evaluate the structural, behavioral, and functional roles of minority nodes using topological metrics, opinion change patterns, and model performance indicators.

H1: Structural Centrality Hypothesis. We find strong support for H1. As shown in Table 1, demographic minority nodes—particularly those defined by parental income, education, and religion—consistently exhibit significantly higher centrality values (degree, betweenness, closeness) compared to their majority counterparts ($p < 0.001$ across metrics). For example, low-income students exhibit a mean degree of 9.01 versus 6.50 for non-minorities.

In contrast, the behavioral minority group (Facebook privacy sensitivity) does not follow this trend. Although their eigenvector centrality is slightly elevated, they display lower degree and betweenness centrality ($p < 0.001$), suggesting more peripheral or localized influence. This contrast helps isolate structural centrality as a feature specific to demographic marginality rather than mere statistical rarity.

Table 1: Centrality differences between minority and majority groups

Group	Metric	p-value	Minority Avg.	Majority Avg.
Privacy	Betweenness	< 0.001	0.0125	0.0149
	Degree	0.070	7.123	7.635
	Eigenvector	< 0.001	0.0624	0.0428
Parental Income	Betweenness	< 0.001	0.0162	0.0129
	Degree	< 0.001	9.012	6.503
	Closeness	< 0.001	0.334	0.309
	Eigenvector	0.241	0.0571	0.0600
Parental Education	Betweenness	< 0.001	0.0178	0.0121
	Degree	< 0.001	9.183	6.611
	Closeness	< 0.001	0.337	0.311
	Eigenvector	< 0.001	0.0705	0.0412
Parental Religion	Betweenness	< 0.001	0.0169	0.0120
	Degree	< 0.001	8.765	6.488
	Closeness	< 0.001	0.335	0.308
	Eigenvector	< 0.001	0.0729	0.0400

H2: Behavioral Adaptability Hypothesis. Demographic minority nodes also exhibit greater opinion fluidity over time, supporting H2. Longitudinal survey analysis shows that minority participants shift positions more often between survey waves, especially on polarized issues like marijuana legalization and equal rights. These changes suggest heightened adaptability or norm sensitivity, in contrast to the stability of majority-group members, whose opinions remain more consistent across time.

H3: Structural Bridging Hypothesis. Minority nodes act as demographic bridges within the network, validating H3. On average, they form 17.2% more cross-group ties than majority nodes ($p < 0.01$), indicating a strong tendency toward demographic boundary spanning. This is further confirmed by lower intragroup clustering coefficients, suggesting minorities serve as connectors between otherwise segregated subgraphs—a classic feature of structural brokerage.

H4: Causal Performance Hypothesis. The exclusion of demographic minorities significantly harms model performance, supporting H4. Across all simulations, removing these nodes leads to a 29% drop in F1 score—from 0.45 ± 0.11 to 0.32 ± 0.07—despite topological and temporal structure being preserved (see Figure 3 and Table 2). The most pronounced degradation occurs in the loss of polarization

Table 2: Key performance metrics comparison

Metric	Original Networks	Minority-Excluded Networks
Mean F1-Score	0.45 ± 0.11	0.32 ± 0.07
Optimal Δ Range	0.3–0.4	N/A (Flat response)
Performance Drop	–	29% (relative)

sensitivity: original networks peak at $\Delta = 0.2$–0.3 (F1 ≈ 0.58), while minority-removed networks yield flattened, non-discriminative responses.

Robustness and Structural Controls. To rule out confounding effects due to structural changes, we performed KS tests on core network properties post-rewiring. Degree distribution ($p = 0.82$), clustering coefficient ($p = 0.79$), and modularity ($p = 0.12$) were preserved within a 15% tolerance, confirming that performance degradation is not due to structural noise but the absence of functionally embedded minority nodes.

Conclusion. These results affirm all four hypotheses (H1–H4), providing causal evidence that demographic minorities are structurally central, behaviorally flexible, and functionally indispensable in sustaining opinion diffusion fidelity. The contrast with the Facebook privacy comparator highlights that not all rare groups exert equal influence—structural marginality, not numerical rarity, underpins minority impact.

Summary and Future Work

This study provides causal evidence that demographic minority nodes are structurally and behaviorally central to the functioning of opinion diffusion systems. Far from being marginal, these nodes occupy pivotal positions in the network and play essential roles in maintaining predictive fidelity. Our findings support the conclusion that their exclusion degrades both the structural coherence and behavioral realism of diffusion models.

We find that minority nodes:

- Hold structurally central and influential positions, bridging disconnected communities and maintaining global connectivity.
- Show greater behavioral adaptability over time, especially on polarized issues.
- Are essential to model fidelity: their removal causes a 29% drop in F1 and eliminates polarization sensitivity, even under preserved topology.

These effects persist across attribute types and are not explained by rarity alone. A behavioral comparator group showed no comparable influence.

Taken together, these results suggest that structural marginality—not merely statistical rarity—underpins the outsize influence of minority groups. Crucially, the inclusion of a behavioral comparator (Facebook privacy preferences) confirms that not all numerically small groups exert similar effects.

Theoretical Implications

Our findings contribute to resolving the long-standing Granovetter-Centola paradox, which highlights two seemingly contradictory mechanisms of diffusion. Granovetter (Granovetter 1973) argued that weak ties—those connecting individuals across community boundaries—enable broad dissemination of novel information. Centola (Centola 2010), by contrast, emphasized that behavioral contagion—particularly for complex behaviors—requires dense clustering and repeated social reinforcement.

While these perspectives appear at odds, our study suggests they are not mutually exclusive when examined through the role of demographically marginal nodes. Minority actors in our networks function as hybrid bridges—individuals embedded in tight-knit communities who also span demographic boundaries. These dual embedded positions enable them to meet the structural prerequisites for both local reinforcement and long-range diffusion.

Rather than viewing tie strength and clustering as incompatible, our results show that diffusion is optimized when certain structurally marginal nodes bridge clustered communities without sacrificing their embeddedness. In this sense, the paradox dissolves: it is not simply the presence of weak or strong ties that matters, but who creates them and where they sit in the network's demographic architecture. Minority nodes, by virtue of their cross-cutting embeddedness, reconcile these competing mechanisms and advance our understanding of how diversity drives network-level diffusion.

CoDiNG Model Enhancement: Structural Awareness

Our findings suggest two extensions to improve the realism and fairness of the CoDiNG model:

1. **Centrality-Weighted Influence Updates**
   The current model assumes equal influence from all neighbors. To reflect the asymmetric role of central nodes—often minorities—we propose weighting peer influence by centrality (e.g., degree, betweenness):
   $$O_j^{\text{new}} = \frac{\sum_{i \in N(j)} w_i \cdot O_i}{\sum_{i \in N(j)} w_i}$$
   This ensures that highly central agents exert more influence, aligning with their observed roles as structural brokers.
2. **Stability-Aware Opinion Retention**
   To account for individual differences in opinion stability, we introduce a stubbornness coefficient $s_i \in [0, 1]$:
   $$O_i^{\text{new}} = (1 - s_i) \cdot O_i^{\text{peer}} + s_i \cdot O_i^{\text{old}}$$
   Lower s_i reflects higher susceptibility to change, typical of minority nodes. This formulation supports simulating tipping effects where stable minorities shift group consensus (Cialdini 1984).

These modifications enable CoDiNG to model both structural asymmetries and behavioral heterogeneity—critical for fairness-aware simulations. While applicable to other opinion models, they increase complexity due to reliance on network-level metrics.

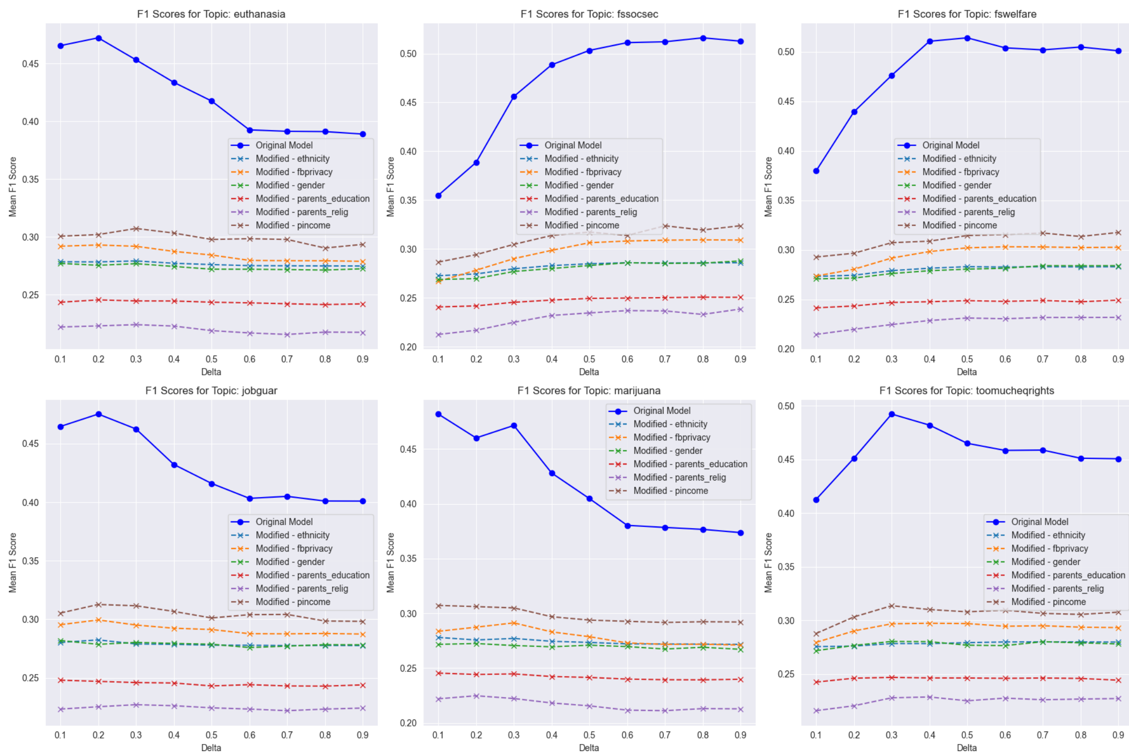

Figure 3: Comparison of F1-scores between original networks and minority-excluded networks across different Δ parameter values

Limitations

This study is based solely on the NetSense dataset, which, while rich in longitudinal detail, is limited to 203 students at a single university over six semesters. This restricts generalizability to broader or more diverse populations.

Additionally, datasets combining high-resolution communication logs and repeated opinion surveys are extremely rare. This limits the feasibility of replication in larger or more heterogeneous settings—an issue of data availability rather than methodological design.

Future Work

With respect to future directions of the research apart from incorporating and evalutating modifications to opinion models proposed above, we propose the following:

1. **Dynamic Minority Identification**
   Replace static thresholds with temporal centrality measures (Borgatti 2005) using NetSense longitudinal data.
2. **Utilizing different datasets**
   Replicate findings using NetHealth dataset in order to increase the sample space of the experiment. (Purta et al. 2016).
3. **Topological Redundancy Assessment**
   Evaluate whether there are redundant minority bridges or if key individuals are uniquely irreplaceable using node removal cascades and centrality-based resilience simulations.

Our results challenge the assumed fairness-accuracy trade-off (Mehrabi et al. 2021), demonstrating that demographic inclusion enhances model fidelity through network structural optimization.

Ethical Statement

This work promotes fairness-aware modeling by exposing structural biases in opinion diffusion networks that emerge from demographic under representation. Our method does not infer protected attributes nor engage in profiling; rather, it simulates causal effects of demographic group presence. All analyses use publicly available, anonymized data under IRB approval. By highlighting the modeling importance of minority inclusion, this work contributes to ethical AI system design.

Acknowledgments

This work was partially funded by the National Science Centre, Poland, grant number 2021/41/B/HS6/02798.

References

Aral, S.; and Walker, D. 2012. Identifying influential and susceptible members of social networks. *Science*, 337(6092): 337–341.

Bass, F. M. 1969. A new product growth for model consumer durables. *Management science*, 15(5): 215–227.

Borgatti, S. P. 2005. Centrality and network flow. *Social networks*, 27(1): 55–71.

Carrington, P. J.; Scott, J.; and Wasserman, S. 2005. *Models and methods in social network analysis*, volume 28. Cambridge university press.

Centola, D. 2010. The spread of behavior in an online social network experiment. *science*, 329(5996): 1194–1197.

Cialdini, R. B. 1984. Influence: The psychology of persuasion. *placeholder*.

Deffuant, G.; Neau, D.; Amblard, F.; and Weisbuch, G. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(01n04): 87–98.

DeGroot, M. H. 1974. Reaching a consensus. *Journal of the American Statistical association*, 69(345): 118–121.

Fan, K.; and Pedrycz, W. 2016. Opinion evolution influenced by informed agents. *Physica A: Statistical Mechanics and Its Applications*, 462: 431–441.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12: 211–223.

Granovetter, M. S. 1973. The strength of weak ties. *American journal of sociology*, 78(6): 1360–1380.

Hofstra, B.; Corten, R.; van Tubergen, F.; and Ellison, N. B. 2017. Sources of Segregation in Social Networks: A Novel Approach Using Facebook. *American Sociological Review*, 82(3): 625–656.

Holme, P.; and Saramäki, J. 2012. Temporal networks. *Physics Reports*, 519(3): 97–125.

Ising, E. 1925. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1): 253–258.

Karimi, F.; Oliveira, M.; and Strohmaier, M. 2022. Minorities in networks and algorithms. *arXiv preprint arXiv:2206.07113*.

Lu, Q.; Korniss, G.; and Szymanski, B. K. 2009. The naming game in social networks: community formation and consensus engineering. *Journal of Economic Interaction and Coordination*, 4(2): 221–235.

Martins, A. C. 2008. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(04): 617–624.

Maslov, S.; and Sneppen, K. 2002. Specificity and stability in topology of protein networks. *Science*, 296(5569): 910–913.

Masuda, N.; and Holme, P. 2020. Guide to temporal networks. *World Scientific*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1): 415–444.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.

Michalski, R.; Szymanski, B. K.; Kazienko, P.; Lebiere, C.; Lizardo, O.; and Kulisiewicz, M. 2021. Social networks through the prism of cognition. *Complexity*, 2021(1): 4963903.

Moran, P. A. 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2): 17–23.

Newman, M. E. J. 2003. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256.

Newman, M. E. J.; and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E*, 69(2): 026113.

Nurek, M.; Kołaczek, J.; Michalski, R.; Szymański, B. K.; and Lizardo, O. 2024. CoDiNG–Naming Game with Continuous Latent Opinions of Individual Agents. *arXiv preprint arXiv:2406.19204*.

Purta, R.; Mattingly, S.; Song, L.; Lizardo, O.; Hachen, D.; Poellabauer, C.; and Striegel, A. 2016. Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, ISWC '16, 28–35. New York, NY, USA: Association for Computing Machinery. ISBN 9781450344609.

Rainer, H.; and Krause, U. 2002. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3): 1–30.

Saxena, A.; Fletcher, G.; and Pechenizkiy, M. 2024. Fairsna: Algorithmic fairness in social network analysis. *ACM Computing Surveys*, 56(8): 1–45.

Striegel, A.; Liu, S.; Meng, L.; Poellabauer, C.; Hachen, D.; and Lizardo, O. 2013. Lessons learned from the netsense smartphone study. *ACM SIGCOMM Computer Communication Review*, 43(4): 51–56.

Sznajd-Weron, K.; and Sznajd, J. 2000. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(06): 1157–1165.

Valente, T. W. 1996. Social network thresholds in the diffusion of innovations. *Social networks*, 18(1): 69–89.

Watts, D. J.; and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684): 440–442.

Zhan, M.; Kou, G.; Dong, Y.; Chiclana, F.; and Herrera-Viedma, E. 2021. Bounded confidence evolution of opinions and actions in social networks. *IEEE Transactions on Cybernetics*, 52(7): 7017–7028.