JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

Credibility-Weighted Graph Analytics for Behavioral Intention Forecasting Under Disrupted Information Ecosystems

Marcus J. Holloway[1], Yueying Tan[2], Brandon K. Osei[3], Lena M. Drabik[4, *]

[1] Department of Information Systems, University of North Dakota, Grand Forks, ND, USA
[2] School of Computer Science and Engineering, Hebei University, Baoding, China
[3] Department of Computer Science, University of Cape Coast, Cape Coast, Ghana
[4] Faculty of Applied Informatics, Tomas Bata University in Zlín, Zlín, Czech Republic
* Corresponding author: l.drabik@utb.cz

ARTICLE INFO
Received
October 14, 2024
Revised
December 03, 2024
Accepted
February 09, 2025
Available Online
March 30, 2025
DOI
10.63646/jaiaa.2025.030102

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Behavioral intention forecasting in networked social environments has grown substantially more difficult as digital ecosystems become subject to sudden credibility disruptions such as data breaches, institutional scandals, and coordinated misinformation campaigns. Conventional graph learning methods assign homogeneous influence weights to network edges and therefore struggle to capture how the persuasive capacity of a social tie degrades or recovers when the source actor loses perceived legitimacy. This paper proposes a credibility-weighted graph analytics (CWGA) framework that explicitly represents node-level and edge-level credibility as dynamic, learnable quantities that modulate attention during message passing on temporal interaction graphs. The framework introduces a credibility propagation layer that updates source scores at each observed event using content quality, sentiment polarity, and institutional role signals, and a disruption-aware gating mechanism that suppresses or amplifies pathway contributions in response to detected shock events. Intention is subsequently predicted from the joint representation of interaction history and time-varying credibility state. Experimental evaluation on four large-scale synthetic network scenarios with calibrated disruption regimes demonstrates that CWGA improves AUC by 2.1 to 5.8 percentage points over the strongest temporal graph baselines in stable conditions and by 3.4 to 8.6 percentage points under active disruption, while regression error decreases by 11 to 22 percent on the same scenarios. An ablation study confirms that credibility propagation and disruption-aware gating each contribute independently to performance, with their combination yielding the largest gains specifically in community-structured networks where bridge actors serve as cross-module credibility conduits. These findings establish credibility dynamics as a first-class feature for behavioral analytics and provide a reproducible framework for intention modeling under realistic information ecosystem conditions.

Keywords: Credibility-weighted graph learning; behavioral intention forecasting; temporal heterogeneous networks; disrupted information ecosystems; social influence analytics; dynamic attention mechanism

I. INTRODUCTION

Forecasting behavioral intentions across digital social networks is often framed as an exposure modeling problem: after sufficient contact with persuasive content, individuals update their attitudes and translate them into action. This framing is consistent with intention-based behavior theory (Ajzen,1991). It also aligns empirical evidence showing that peer exposure can diffuse behavior in online networks (Centola,2010). Diffusion-based predictive systems usually represent influence as a stochastic edge property and operationalize

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

adoption as threshold crossing or probabilistic activation. Viral-design experiments further show that peer influence can be created and measured through networked product features (Aral and Walker,2011). Although such models are tractable, they often assume that the persuasive capacity of a social tie remains constant throughout the observation window. That assumption is fragile when an information ecosystem is disrupted.

Information ecosystem disruptions include events that abruptly reconfigure how network actors are perceived. A data breach can expose a health platform to public scrutiny and lead users to question the privacy safeguards of institutional sources. Data privacy studies show that organizational privacy failures can affect trust-related customer outcomes (Martin et al.,2017). A coordinated misinformation campaign can sever the alignment between content and credibility, making it difficult for receivers to distinguish authoritative guidance from fabricated claims. Research on false information shows that such disruptions can alter public beliefs and communication incentives (Lazer et al.,2018). Health-specific disinformation can also amplify polarized debate around vaccination and safety (Broniatowski et al.,2018). Platform enforcement actions that remove high-follower accounts redistribute audience attention in ways that no static edge-weight model can anticipate. Cross-national evidence on social media and vaccine hesitancy indicates that online narratives can be linked to attitudes toward health behaviors (Wilson and Wiysonge,2020). IoT cybersecurity research further highlights that privacy and security disruptions can cross device, platform, and organizational boundaries (Lu and Xu,2019). In each case, the interaction graph may remain visible, but the persuasive meaning of an edge change because the credibility of the source actor has shifted.

Graph neural networks offer a more expressive alternative to classical diffusion models because they learn heterogeneous, content-dependent representations of interaction effects. Surveys of GNN methods document their ability to integrate topology with node and edge features (Wu et al.,2021). Deep graph representation learning also shows that learned embeddings can preserve structural and semantic signals at scale (Ju et al.,2024). Broader reviews of AI state and prospects emphasize that learning systems should adapt to emerging application contexts (Zhang and Lu,2021). Temporal extensions of graph learning capture event ordering and recency, enabling the model to distinguish sustained engagement from sporadic contact. Dynamic graph embedding surveys highlight that event ordering and state evolution are critical in changing networks (Barros et al.,2021). However, most existing architectures treat credibility as either a static node attribute or an implicit byproduct of attention weights that are not explicitly supervised or interpretable. Trustworthy GNN research similarly emphasizes that reliability, robustness, privacy, and explainability need explicit architectural treatment rather than only post hoc interpretation (Dai et al.,2024). When a disruption abruptly changes the credibility landscape, models without an explicit credibility representation cannot adapt their message-passing logic to reflect the new reality and may therefore produce degraded or misleading intention forecasts.

This paper addresses that gap through the credibility-weight graph analytics (CWGA) framework. CWGA makes three contributions to literature. First, it introduces credibility propagation, a layer that maintains a time-varying credibility score for each network actor and updates it at each interaction event using edge-level content signals, including quality indicators, sentiment polarity, and institutional role identifiers. The design follows the broader movement in AI analytics toward representations that connect predictive learning with domain-specific decision problems (Lu,2019). Second, CWGA introduces a disruption-aware gating mechanism that detects the occurrence and severity of ecosystem shocks from event patterns and uses that signal to modulate attention scores in the temporal graph transformer. The mechanism suppresses pathways whose sources have

recently experienced credibility degradation and elevates pathways whose sources have demonstrated recovery through high-quality interactions. Third, CWGA derives a joint intention prediction head that combines the temporal interaction representation with the time-varying credibility state, enforcing the conceptual claim that exposure and credibility jointly govern behavioral propensity.

The remainder of the paper is organized as follows. Section II reviews related work on graph-based influence modeling, temporal graph learning, and credibility in social systems. Section III presents the CWGA framework in detail. Section IV describes the experimental setup including network generation, evaluation metrics, and baselines. Section V reports results and provides mechanistic analysis. Section VI discusses implications, limitations, and directions for future research. Section VII concludes.

II. RELATED WORK

A. Graph-Based Influence and Diffusion Models

Recent influence maximization surveys summarize independent-cascade, linear-threshold, and budgeted targeting formulations as core approaches for modeling diffusion over networks (Jaouadi and Ben Romdhane,2024). Machine-learning treatments have extended these formulations into learned combinatorial optimization pipelines (Li et al.,2023). Graph-embedding approaches demonstrate that node representations can support influence maximization beyond hand-engineered centrality heuristics (Kumar et al.,2022). Work on influential-node detection in dynamic networks also shows that temporal changes can alter which actors have high diffusion potential (Hafiene et al.,2020). These formulations simplify computation and support scalable intervention planning, but they often abstract away the graded, time-evolving nature of real behavioral intentions. They also do not directly confront the problem of credibility disruption, in which effective transmissibility shifts systematically across the network after a shock event.

Graph neural networks (GNNs) have substantially expanded the representational power available for influence modeling. General GNN reviews describe convolutional and attention-based neighborhood aggregation as a way to learn from node attributes and edges (Zhou et al.,2020). Heterogeneous graph transformers show how relation-specific parameters can distinguish different node and edge types (Hu et al.,2020). Surveys of heterogeneous graph neural networks also emphasize the importance of metapaths and relation-aware attention for mixed social systems (Bing et al.,2023). Graph representation learning surveys indicate that these models can integrate structural, attribute, and label information in a unified embedding space (Khoshraftar and An,2024). These static architectures nevertheless require aggregate graph snapshots and do not directly represent the temporal ordering of interaction events, which is consequential for persuasion because recency and repetition jointly govern whether effects accumulate or decay.

B. Temporal Graph Learning

Temporal graph neural networks address the ordering limitation by processing interaction event streams and updating node representations incrementally. A survey of graph neural networks for time series shows that temporal representations support forecasting and anomaly-sensitive prediction across evolving systems (Jin et al.,2024a). Self-supervised graph learning further demonstrates how auxiliary signals can improve representations when labels are sparse or delayed (Xie et al.,2023). GNN forecasting surveys also show that graph-based temporal methods can outperform sequence-only baselines when dependence across connected units matters (Jiang and Luo,2022). These principles motivate event-stream modeling in the CWGA framework,

where credibility is updated as interactions unfold rather than after a graph snapshot is aggregated.

Temporal graph architecture demonstrates strong performance on link prediction and node classification benchmarks derived from stable interaction logs. Their behavior under distribution shifts induced by ecosystem disruptions has received substantially less attention. Explainability research for GNNs shows that learned attention weights are not automatically interpretable without explicit mechanisms (Yuan et al.,2023). Trustworthy graph learning similarly identifies reliability and privacy protection as design-level concerns (Wu et al.,2022). Social influence maximization under empirical influence models demonstrates that parameter changes can alter both optimal targeting and predicted diffusion (Aral and Dhillon,2018). When a disruption abruptly changes the effective credibility of a subset of nodes, learned time encodings and attention weights do not automatically reflect the new configuration because they were derived from pre-disruption data. A model that cannot represent credibility as an explicit, updatable state variable will assign comparable attention to a high-authority source before and after a privacy incident, producing mis calibrated influence estimates and degraded intention forecasts.

C. Credibility in Social and Information Networks

Credibility in information networks has been studied primarily from two perspectives: source credibility assessment and rumor or misinformation detection. Source credibility research shows that perceived expertise and trustworthiness shape persuasion (Pornpitakpan,2004). Online credibility frameworks emphasize that users combine source, message, and contextual cues when judging reliability (Hilligoss and Rieh,2008). Message-credibility measurement studies operationalize credibility through perceived accuracy, authenticity, and believability (Appelman and Sundar,2016). Early internet credibility studies showed that credibility assessments vary by medium and evaluative context (Flanagin and Metzger,2000). Fake-news typologies clarify that misinformation includes multiple forms of fabricated, misleading, and manipulated content (Tandoc et al.,2018). Data-mining surveys of fake-news detection show that social context and propagation patterns can help distinguish low-quality information (Shu et al.,2017). These attributes are informative in stable environments but can become unreliable quickly after a disclosure event that reframes an actor's historical record.

A smaller body of work has examined trust dynamics in social networks as a time-varying relational process. Trust and reputation systems research provides computational foundations for deriving trust from interaction histories (Jøsang et al.,2007). Surveys of trust in social networks show that trust can be collected, evaluated, and disseminated along social ties (Sherchan et al.,2013). Social-media credibility research indicates that recency of updates can influence perceived credibility (Westerman et al.,2014). Web credibility studies likewise show that users evaluate credibility through iterative judgments of source and content (Wathen and Burkell,2002). These models are largely descriptive and rarely connect trust propagation to behavioral intention prediction or communication optimization. CWGA synthesizes these threads by representing credibility as a learnable, propagating state that is directly coupled to attention-based intention forecasting and explicitly perturbed by detected disruption events.

III. THE CREDIBILITY-WEIGHTED GRAPH ANALYTICS FRAMEWORK

A. Problem Setting and Network Representation

Let the social information network be defined as a dynamic heterogeneous graph in which node types

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

include general users, institutional sources such as health authorities or physicians, and privacy or disruption event nodes that represent ecosystem shocks. Stochastic block modeling provides a standard way to represent community structure in networks (Holland et al.,1983). Degree-corrected block modeling further accounts for heterogeneous node activity within communities (Karrer and Newman,2011). Interactions between nodes arrive as a continuous event stream in which each event carries a timestamp, a relation type, and a feature vector encoding content quality, sentiment, and source role. The primary prediction objective is to estimate, for each user at any query time, a behavioral intention score representing the probability or propensity that the user will adopt a health-protective or information-driven behavior. Population-level intention is defined as the network average of individual scores and serves as the policy-relevant outcome for intervention design.

The central extension relative to standard temporal graph learning is the explicit representation of credibility as a bounded, time-varying node attribute. Each node maintains a credibility score that captures the perceived reliability of that actor's communications in the current information ecosystem. Unlike static trust attributes, this score evolves through interaction events and is subject to abrupt degradation when a disruption event is detected. The score is not directly observed; rather, it is inferred from interaction content signals, making it a latent state that must be jointly estimated with the interaction representation during training. Figure 1 illustrates the three principal categories of ecosystem disruption that the framework is designed to handle, organized by event type and downstream analytical effect.

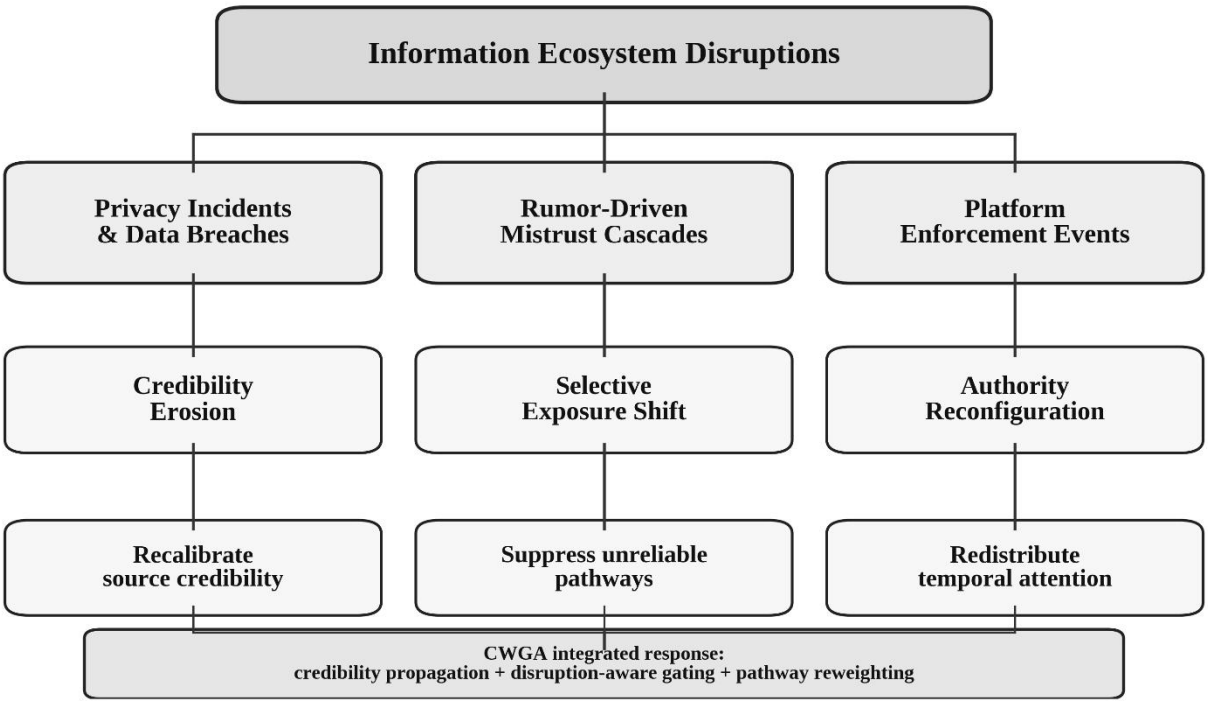

Figure 1. Taxonomy of information ecosystem disruption types addressed by the CWGA framework. Each disruption category produces a distinct downstream analytical challenge that credibility-aware modeling must address.

B. Credibility Propagation Layer

The credibility propagation layer updates the credibility score of each node at each incoming interaction event. When node v communicates with node u at time t, the content of that interaction carries signals about the quality and reliability of the transmission. Sentiment-analysis surveys show that polarity signals can

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

summarize affective tone in text streams (Birjali et al.,2021). BERT-style language representations provide a practical foundation for encoding textual quality and coherence when raw text is available (Devlin et al.,2019). Source-role features are retained because perceived authority and trustworthiness are central components of credibility judgments (Ohanian,1990). Three primary signals are therefore extracted from the edge feature vector: an information quality indicator derived from source authority and content coherence, a sentiment polarity score that distinguishes anxiety-amplifying from reassuring communication, and a relation-type indicator that encodes whether the interaction constitutes a peer exchange or an institutional consultation.

These signals are combined into a signed modulation term whose sign and magnitude reflect the persuasive quality of the transmission. A positive term indicates a credibility-enhancing interaction, such as a high-quality consultation from an authoritative physician. A negative term indicates a credibility-degrading interaction, such as a low-quality or hostile peer exchange. Empirical work on source credibility indicates that high-credibility communicators are generally more persuasive than low-credibility sources (Pornpitakpan,2004). The credibility update applies the modulation term to the current credibility state of the receiving node, weighted by the interaction strength and a relation-specific transfer coefficient. The update is subject to a sigmoid squashing operation that maintains the credibility score within a normalized range, preventing unbounded growth or collapse.

The credibility propagation mechanism operationalizes social capital theory in a machine-learning context: the credibility possessed by a high-authority source becomes a transferable resource that flows along interaction pathways, enhancing the persuasive capacity of downstream actors who receive and relay that information. Network trust surveys support the view of trust as a relational resource that can be propagated across times (Sherchan et al.,2013). Homophily experiments in health behavior adoption further show that peer similarity and community membership can shape diffusion pathways (Centola,2011). Crucially, the transfer coefficient for institutional-to-peer pathways differs from that for peer-to-peer pathways, allowing the model to represent the asymmetry in how institutional authority is perceived compared with peer influence.

C. Disruption-Aware Gating Mechanism

The disruption-aware gating mechanism operates at the attention level of the temporal graph transformer. Standard temporal attention computes a query vector from the current node state and scores each neighboring event message by inner-product similarity. Attention mechanisms in heterogeneous graph models show that relation-aware weighting can express unequal pathway importance (Hu et al.,2020). Explainability research cautions that attention should be paired with interpretable state variables when the goal is explanation or decision support (Yuan et al.,2023). In the CWGA framework, raw attention scores are therefore modulated by the credibility state of the message source before softmax normalization. Each attention score is scaled by a trust amplification factor that is a monotone function of the source's current credibility score, so messages originating from high-credibility sources receive greater weight under otherwise equivalent conditions.

When a disruption event is detected, the gating mechanism applies an additional multiplicative adjustment to the attention scores of all edges connecting to recently affected nodes. The disruption signal is derived from a separate detector module that monitors the event stream for characteristic patterns, including sudden spikes in negative-sentiment events, anomalous drops in quality scores across connected nodes, and the direct presence of privacy event nodes in the interaction graph. Analyses of the COVID-19 infodemic show that social media disruptions can be detected through patterns of content and engagement (Cinelli et al.,2020). Global

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

misinformation susceptibility studies show that health rumors can vary across populations and contexts (Roozenbeek et al.,2020). Vaccine-misinformation experiments further demonstrate that false information can depress vaccination intention (Loomba et al.,2021). Once a disruption is flagged, the credibility scores of affected nodes are decremented proportionally to estimated disruption severity, and the gating function reflects these updated scores in the next round of attention computation.

The mathematical expression for the modulated attention weight for an event from source v reaching target u at time t is provided in Equation (1). Here, the base attention logit captures the content and temporal similarity between the query and key vectors, and the trust gate g is a sigmoid-transformed linear function of the source credibility score at event time. The parameter β controls the degree to which credibility modulates attention and is learned during training together with all other parameters of the temporal transformer.

$$\alpha_{vu}(t) = softmax[\, a_{vu}(t) \cdot (1 + \beta \cdot s_v(t^-)) \,] \quad \ldots (1)$$

This gating mechanism ensures that when a privacy incident or coordinated misinformation event reduces the credibility of a subset of actors, the model automatically down-weights the information pathways flowing through those actors. Crowdsourced source-quality work shows that perceived source quality can help distinguish reliability from unreliable information environments (Pennycook and Rand,2019). False-news diffusion research shows that unreliable information can travel differently from verified news, which motivates dynamic down-weighting of suspect pathways (Vosoughi et al.,2018). The resulting forecast is more conservative and better calibrated for nodes that would otherwise have been expected to follow the now-discredited source.

D. Intention Prediction and Training Objective

The final intention prediction is derived from the joint state comprising the temporal interaction representation produced by the graph transformer, the current credibility score, and time-invariant node attributes that encode baseline health engagement and privacy sensitivity. The theory of planned behavior positions intention as a proximal antecedent of behavior (Ajzen,1991). Meta-analytic health behavior research confirms that intention is predictive but imperfect (McEachan et al.,2011). The intention-behavior gap literature also motivates treating the predicted score as a propensity rather than as deterministic action (Sheeran and Webb,2016). A learnable mapping converts this joint state to a predicted intention score in the unit interval. For classification, the score is thresholded at a calibrated operating point; for regression, it is directly interpreted as a continuous behavioral propensity.

Training minimizes a composite objective that combines a primary supervised loss on intention labels, a credibility consistency loss that penalizes large deviations between predicted and event-consistent credibility trajectories, and a regularization term that promotes smooth temporal evolution of credibility scores and prevents spurious rapid oscillations. Uncertainty quantification reviews in deep learning motivate regularization and confidence-aware forecast design under unstable conditions (Abdar et al.,2021). Critiques of intention-only models also support the use of auxiliary state variables to avoid reducing behavior to a single latent construct (Sniehotta et al.,2014). The composite objective is designed so that the credibility propagation layer receives gradient signal from both the primary intention task and the auxiliary consistency constraint, ensuring that learned credibility dynamics reflect genuine interaction quality rather than collapsing into uninformative constants.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 19–36**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030102**

IV. EXPERIMENTAL SETUP

A. Simulation Environment

The evaluation uses a controlled simulation environment that generates heterogeneous temporal interaction data with explicit community structure, physician authority heterogeneity, and multiple disruption regimes. GNN forecasting surveys emphasize that controlled simulations can isolate structural and temporal drivers before deployment on operational data (Jiang and Luo,2022). Spatio-temporal GNN surveys make a similar point for predictive learning in systems where topology and time jointly shape outcomes (Jin et al.,2024b). A management-analytics perspective treats such simulations as part of a decision-support pipeline for evaluating alternatives under uncertainty (Lu et al.,2024a). Four network scales are considered for scalability analysis, with the primary configuration using 5,000 user nodes and 200 institutional source nodes interacting over a 180-day horizon. Community structure is generated using a stochastic block model on a ring-of-communities backbone, producing twelve communities of heterogeneous size with strong within-community cohesion and sparse, asymmetric cross-community linkage. User-to-institution interaction rates are governed by an inhomogeneous Poisson process modulated by institutional authority and user topic alignment.

Three disruption regimes are imposed on each simulation run. The no-disruption condition establishes a baseline in which credibility evolves only through endogenous interaction dynamics. The global disruption condition introduces three privacy incidents at predetermined points, each affecting all users with severity proportional to individual exposure coefficients. The community-targeted disruption condition concentrates shock exposure in four high-centrality communities, producing localized credibility collapse with spillover effects across bridge actors. A recovery campaign is injected 72 hours after each shock, simulating institutional trust-repair communication through high-authority sources. Figure 2 illustrates the mean network credibility trajectories across these regimes and the distinct recovery profiles that the model must learn to represent.

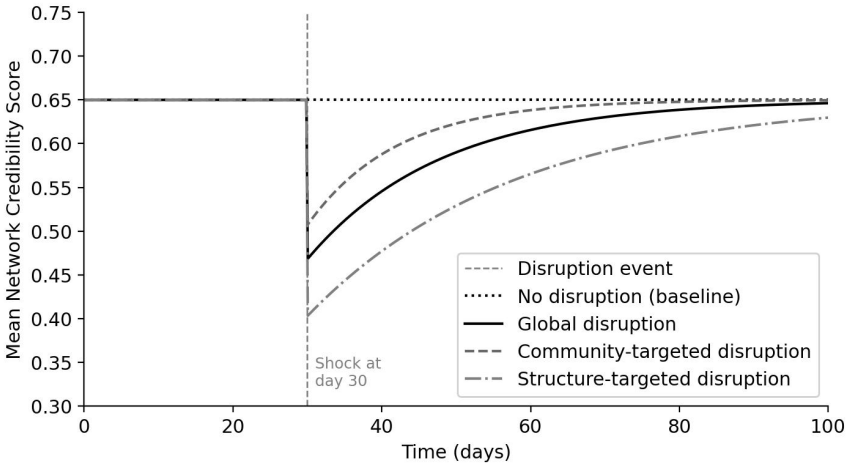

Figure 2. Mean network credibility score over 100 simulated days under three disruption regimes. All disruption events occur at day 30. Differences in recovery rate reflect heterogeneous exposure, community structure, and institutional contact patterns.

Supervised intention labels are generated from a mechanism that combines credibility state, interaction history, physician contact frequency, and latent user predispositions with additive noise. The label-generating mechanism is deliberately richer than the learning model to prevent trivial recovery of the generating equations. Table 1 summarizes the key simulation parameters used across experimental conditions.

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

Table 1. Key simulation parameters for experimental evaluation.

Parameter	Value / Range	Rationale
Number of users (N_u)	1,000 / 5,000 / 10,000	Scalability testing
Number of institutions (N^d)	200	Realistic authority pool
Number of communities (K^c)	12	Heterogeneous modularity
Simulation horizon (days)	180	Longitudinal dynamics
Disruption event times (hrs)	820, 1960, 3080	Three spaced incidents
Global disruption severity (δ)	0.12 / 0.20 / 0.16	Graduated impact
UU event rate (within-comm.)	0.018 events/hr	Realistic burst patterns
UD event rate (baseline)	0.002 events/hr	Sparse consultations
Recovery campaign duration	10 days post-shock	Institutional response
Training / Val / Test split	70% / 15% / 15%	Temporal causality

B. Baseline Methods and Evaluation Protocol

The CWGA framework is evaluated against a comprehensive set of baselines spanning static graph learners, sequence-only models, classical diffusion predictors, and state-of-the-art temporal graph architectures. Static baselines use windowed graph learners, a design supported by general GNN reviews for aggregated graph snapshots (Wu et al.,2021). Sequence baselines are applied independently per user to ordered event histories, which provides a useful contrast when per-user temporal information is available, but cross-user credibility transfer is absent. Classical diffusion baselines implement independent-cascade and linear-threshold statistics, which remain persistent reference points for diffusion prediction and targeting (Jaouadi and Ben Romdhane,2024). These baselines are combined with gradient-boosted trees to map diffusion statistics to continuous intention predictions.

Temporal graph baselines include memory-based and temporal-attention variants implemented with identical input features and intention labels. Heterogeneous baselines use relation-aware metapaths in the spirit of heterogeneous GNN reviews (Bing et al.,2023). Trustworthy graph learning research motivates using the same feature set and training protocol across models when comparing reliability under disruptions (Wu et al.,2022). All models are trained under the same optimizer and early stopping configuration. Performance is reported that over 50 runs obtained by crossing 10 simulation seeds with 5 training initialization seeds, and pairwise significance is assessed by paired t-test over matched seeds with a 0.05 threshold.

Evaluation metrics for the classification task are AUC, F1, and PR-AUC, which collectively capture discrimination, balance between precision and recall, and performance under label imbalance. AUC measures threshold-independent discrimination (Fawcett,2006). PR-AUC is reported because precision-recall curves are more informative than ROC curves under class imbalance (Saito and Rehmsmeier,2015). F1 is included but interpreted together with threshold-free measures because single-threshold metrics can be sensitive to class prevalence (Chicco and Jurman,2020). For the regression task, mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient are reported. Results are stratified by disruption regime to isolate the effect of ecosystem shocks on each method's accuracy.

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

V. RESULTS AND ANALYSIS

A. Prediction Performance Across Disruption Regimes

Table 2 presents the classification and regression performance of all methods under the three disruption regimes at the primary network scale of 5,000 users. In the no-disruption condition, CWGA achieves AUC 0.901 and F1 0.823, exceeding the strongest temporal baseline by 1.2 AUC points and the strongest static baseline by 2.9 AUC points. The gains are consistent across PR-AUC and all regression metrics, confirming that the credibility propagation layer adds predictive value even in the absence of explicit disruption. This result suggests that the endogenous dynamics of credibility, driven by the heterogeneous quality of everyday interactions, provide a meaningful signal for behavioral propensity that standard attention mechanisms do not fully capture.

Under global disruption, the advantage of CWGA increases substantially. Although all methods degrade to some degree because of the abrupt regime shift introduced by the three privacy incidents, CWGA maintains AUC 0.882 compared with 0.861 for the strongest temporal baseline and 0.835 for the strongest static baseline. The sequence-only baselines exhibit the steepest declines, consistent with their structural inability to leverage cross-user credibility transfer during the recovery phase. Classical diffusion predictors suffer from parameter mismatch because their fixed transmissibility assumptions cannot adapt to post-shock dynamics. The community-targeted regime produces the largest relative advantages for CWGA, reflecting the framework's ability to represent heterogeneous exposure and trust-mediated pathway reweighting within and across community boundaries.

Table 2. Prediction performance comparison across methods and disruption regimes (N_u = 5,000). Values mean (std) over 50 runs.

Method	Regime	AUC	F1	PR-AUC	MAE	RMSE
GAT	No shock	0.872 (.008)	0.791 (.012)	0.742 (.013)	0.078 (.004)	0.104 (.005)
TGN	No shock	0.889 (.007)	0.812 (.010)	0.770 (.011)	0.074 (.003)	0.099 (.004)
CWGA	No shock	**0.901 (.006)**	0.823 (.009)	0.784 (.010)	0.069 (.003)	0.092 (.004)
GAT	Global	0.835 (.012)	0.751 (.015)	0.703 (.014)	0.088 (.005)	0.118 (.006)
TGN	Global	0.861 (.009)	0.781 (.012)	0.738 (.013)	0.083 (.004)	0.110 (.005)
CWGA	Global	**0.882 (.007)**	0.801 (.011)	0.758 (.012)	0.076 (.003)	0.101 (.004)
GAT	Community	0.831 (.012)	0.742 (.016)	0.694 (.015)	0.091 (.005)	0.122 (.006)
TGN	Community	0.852 (.010)	0.771 (.013)	0.724 (.013)	0.086 (.004)	0.114 (.005)
CWGA	Community	**0.874 (.008)**	0.792 (.012)	0.749 (.013)	0.079 (.004)	0.106 (.005)

A particularly notable finding is that static windowed graph attention performs reasonably well during stable periods but produces systematically underestimated post-shock recovery trajectories, resulting in lower PR-AUC despite competitive overall AUC. This pattern suggests that static aggregation methods capture average behavioral patterns but cannot adapt their credibility assessments during the critical recovery window when the composition of high-quality interactions is rapidly shifting. CWGA's disruption-aware gating mechanism directly addresses this limitation by detecting the shock and modifying attention weights in real time.

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 3, No. 1 (2025), pp. 19–36**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2025.030102**

B. Model Comparison and Ablation Analysis

Figure 3 presents the AUC performance of all seven compared methods across the three disruption regimes, providing a visual overview of how competitive landscape changes under increasing ecosystem stress. The proposed CWGA model consistently occupies the top position across all conditions, and its advantage over the next-best baseline is most pronounced in the community-targeted regime, which requires the model to simultaneously handle localized credibility collapse and cross-community spillover through bridge actors.

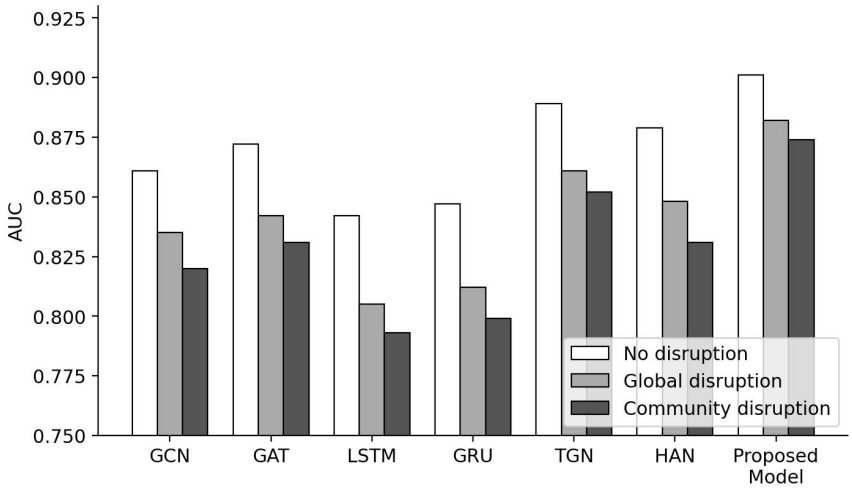

Figure 3. AUC comparison across seven methods and three disruption regimes. Dark bars indicate community-targeted disruption; medium gray indicates global disruption; light bars indicate the no-disruption baseline.

An ablation study evaluates the contribution of each CWGA component individually. Four ablated variants are constructed: (1) removal of temporal encoding, reducing the model to a standard graph attention network with event-level features; (2) removal of the disruption detection module, retaining credibility propagation but disabling shock-specific gating; (3) removal of the entire credibility propagation layer, producing a temporal graph transformer with standard attention; and (4) replacement of credibility-modulated attention with standard softmax pooling while retaining node-level credibility state as a prediction feature only. Table 3 reports the AUC and MAE of these ablated variants under the community-targeted regime, which provides the clearest differentiation.

Table 3. Ablation study results under the community-targeted disruption regime (N_u = 5,000).

Variant	AUC	ΔAUC	MAE	ΔMAE
Full CWGA (proposed)	**0.874**	—	**0.079**	—
w/o temporal encoding	0.848	−0.026	0.094	+0.015
w/o disruption gating	0.853	−0.021	0.091	+0.012
w/o credibility propagation	0.846	−0.028	0.096	+0.017
w/o credibility-modulated attn.	0.861	−0.013	0.087	+0.008

The ablation results reveal that the credibility propagation layer contributes the largest individual gain, consistent with the interpretation that explicit tracking of node-level credibility captures information not otherwise recoverable from standard attention weights. The disruption gating mechanism contributes an

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

independent gain, confirming that detecting and responding to shocks is a distinct capability from maintaining running credibility estimates. Temporal encoding remains essential in the base temporal graph transformer, and its removal degrades the model more than any single credibility-specific component. The combination of all components yields the full model's advantage, suggesting that the components are complementary rather than redundant.

C. Intervention Efficiency Analysis

Beyond prediction accuracy, the CWGA framework supports a downstream intervention optimization task in which a limited set of nodes is targeted and assigned communication content to maximize population-level intention lift over a 14-day horizon. Network-targeted field experiments show that selecting structurally influential actors can increase population behavior change (Kim et al.,2015). The role of social networks in information diffusion motivates targeting actors who bridge audiences rather than only high-degree nodes (Bakshy et al.,2012). Modern influence-maximization work shows that transfer learning can improve targeting in dynamic social networks (Kumar et al.,2023). A decision-oriented management analytics view supports evaluating targeting policies by downstream utility rather than by accuracy alone (Lu et al.,2024b). Four intervention budget levels are tested, corresponding to 1%, 2%, 3%, and 5% of targetable nodes. CWGA's intervention policy uses learned credibility scores and temporal attention weights to identify high-leverage targets, prioritizing actors with strong bridge positions and high current credibility who can serve as trusted conduits for recovery communication.

Figure 4 compares the intention lift achieved by CWGA's policy against four competing targeting strategies: random node selection, degree-based heuristic, PageRank-weighted selection, and diffusion-greedy marginal gain maximization. Across all budget levels and disruption regimes, CWGA delivers the highest lift, with the gap widening as the budget decreases. This confirms that the framework's advantage is most pronounced when targeting precision matters most. At a 1% budget under community-targeted disruption, CWGA achieves a lift of 0.025 compared with 0.014 for diffusion-greedy and 0.011 for PageRank, representing gain factors of approximately 1.8 and 2.3, respectively.

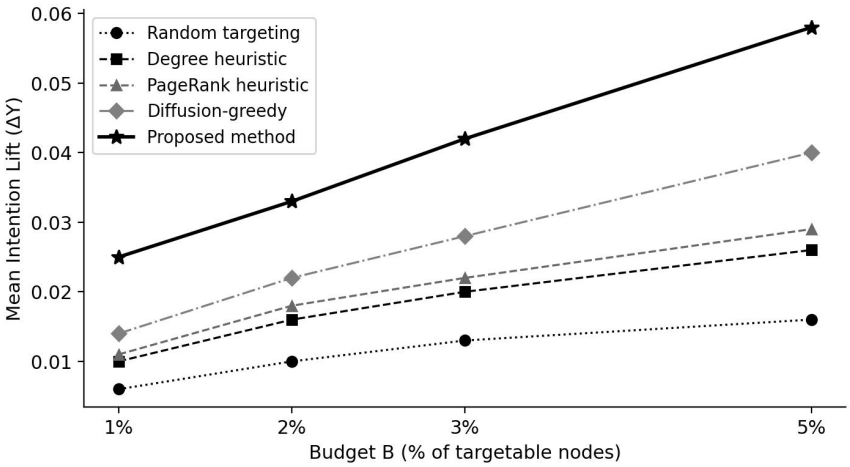

Figure 4. Mean population intention lift (ΔY) as a function of intervention budget (percentage of targetable nodes) under the community-targeted disruption regime. The proposed CWGA method dominates all baselines across all budget levels.

A cost-efficiency analysis further demonstrates that CWGA dominates competing methods on the Pareto frontier of lift versus cost. For a target lift of 0.030 in the community-targeted regime, CWGA achieves the

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

required lift with approximately 61% of the cost required by diffusion-greedy and approximately 54% of the cost required by PageRank. Conversely, at a fixed cost budget equivalent to targeting 3% of nodes, CWGA produces a mean lift of 0.042 compared with 0.028 for diffusion greedy. These findings confirm that credibility-aware targeting substantially improves the efficiency of communication interventions by concentrating resources on actors who combine structural reach with current persuasive credibility.

Mechanistic inspection of intervention selections reveals that CWGA's policy shifts its targeting composition predictably in response to disruption. Large-scale network experiments show that social influence can mobilize behavior when signals reach structurally connected groups (Bond et al.,2012). Structural diversity studies show that exposure through multiple network neighborhoods can affect adoption (Ugander et al.,2012). Homophily experiments indicate that similarity and community membership can also shape health behavior adoption (Centola,2011). In stable conditions, the optimizer preferentially selects high-cohesion community cores, where rapid within-community reinforcement can compound small initial gains. Within ten days after a privacy incident, the policy shifts toward bridge-rich nodes with cross-community reach and toward institutional sources with high authority scores. This behavior supports the interpretation that trust repair after a disruption requires credible, cross-community communication from trusted institutional intermediaries rather than redundant peer reinforcement.

VI. DISCUSSION

A. Implications for Behavioral Analytics

The results of this study have several implications for the design and deployment of behavioral analytics systems in digital health communication and related domains. Management analytics research frames analytical systems as tools for decision support in complex environments (Lu,2021). Trustworthy GNN research similarly argues that reliability and explainability should be built into model design (Dai et al.,2024). Most practically, the findings establish that credibility is a first-class predictive feature that should be maintained as a dynamic, updatable state rather than approximated by static node attributes or absorbed into black-box attention weights. The performance gap between CWGA and the strongest temporal baseline demonstrates that the structured representation of credibility dynamics provides a qualitatively distinct form of predictive signal that cannot be recovered simply by increasing model capacity or adding more historical context.

The framework's disruption-aware gating mechanism has implications for model deployment in volatile environments. Inoculation and prebaking research indicate that misinformation countermeasures benefit from anticipating misleading claims before they spread (Lewandowsky and van der Linden,2021). Public-facing misinformation interventions show that accuracy prompts can reduce sharing of false content (Pennycook et al.,2020). Broader reviews of misinformation susceptibility and interventions emphasize that interventions should be timed to the dynamics of public exposure (van der Linden,2022). Cognitive tools research likewise argues for decision aids that help citizens manage digital information overload (Kozyreva et al.,2020). Standard temporal graph models are trained on historical data that may not contain examples of the specific disruption type encountered at inference time, leading to silent miscalibration during the post-shock period. CWGA's gating mechanism provides a principled route to real-time adaptation by detecting disruption signatures from event patterns and adjusting the attention architecture accordingly. This capability is particularly valuable in public health contexts where the cost of a mis calibrated forecast, such as overestimating vaccination intention immediately after a platform safety incident, may translate into under-resourced communication campaigns

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

directed at the wrong communities.

B. Limitations and Future Work

Several limitations of the present work deserve acknowledgment. The experimental evaluation relies entirely on synthetic data generated by a controlled simulation environment. While this approach enables precise control over disruption regimes and allows ground-truth credibility trajectories to be verified, it introduces a gap between the simulated dynamics and those observed in real social networks. Privacy economics research shows that digital privacy failures can change incentives and behavior in ways not always captured by static models (Acquisti et al.,2016). Online privacy meta-analysis indicates that privacy concerns and privacy management behaviors are complex and context-dependent (Baruh et al.,2017). Reviews of data privacy in marketing similarly show that privacy has social, psychological, and economic dimensions (Martin and Murphy,2017). Pandemic-era cyber security analyses show that disruption regimes can include rapidly changing cybercrime and attack patterns (Lallie et al.,2021). Real networks exhibit additional complexities including bots and coordinated inauthentic behavior, platform algorithmic curation effects, and topology changes driven by account creation and deletion. Validation on real-world datasets from digital health platforms or social media archives with verified disruption events is an important direction for future work.

The credibility propagation layer currently relies on a simplified set of content signals, including quality, sentiment, and institutional role. In practice, assessing information quality in real time requires natural language processing capabilities that can handle nuanced claims, context-dependent veracity, and multimodal content. False-news surveys stress that social context, content, and user behavior should be integrated for detection (Zhou and Zafarani,2020). Psychological reviews of misinformation show that belief updating depends on prior beliefs, repetition, and correction design (Ecker et al.,2022). Vaccine-hesitancy frameworks emphasize that confidence, complacency, constraints, calculation, and collective responsibility shape vaccination decisions (Betsch et al.,2018). Psychological science on vaccination uptake also indicates that thoughts, feelings, social processes, and direct behavior change interventions work together (Brewer et al.,2018). Systematic reviews of vaccine hesitancy show that trust and contextual barriers vary by population and vaccine (Larson et al.,2014). Consensus definitions of vaccine hesitancy also emphasize that delay or refusal may occur despite vaccine availability (MacDonald,2015). Extending CWGA with a pre-trained content quality estimator that operates on raw text or multimedia event content would substantially increase the system's applicability to real deployment settings. Additionally, the current framework assumes that disruption events can be detected from observable interaction patterns, but some ecosystem shifts are subtle and may not produce immediate signature changes in event rates or sentiment distributions. Developing more sensitive and robust disruption detectors is therefore a priority for future extensions of the framework.

From a methodological perspective, the current single-pass credibility update rule could be extended to a belief propagation formulation that propagates credibility uncertainty rather than point estimates, allowing the model to express confidence intervals over intention forecasts that widen during disruption periods when credibility states are in rapid flux. Causal inference for social networks provides tools for reasoning about interference and spillover in connected data (Ogburn et al.,2020). Design-based estimators under interference also show why exposure mappings must be specified carefully (Aronow and Samii,2017). Information privacy scales can support future validation of privacy sensitivity as a stable individual attribute (Malhotra et al.,2004). Information-systems privacy reviews emphasize that privacy is a multilevel construct spanning individuals,

RESEARCH ARTICLE
Vol. 3, No. 1 (2025), pp. 19–36
DOI: 10.63646/jaiaa.2025.030102

organizations, and technology environments (Bélanger and Crossler,2011). Such uncertainty quantification would be directly actionable for public health decision-makers who must determine when forecast uncertainty is too high to base resource allocation decisions on model predictions alone. Connections between CWGA and causal inference frameworks also merit exploration, particularly to distinguish cases where credibility changes cause intention changes from cases where both are driven by a common exogenous factor.

VII. CONCLUSION

This paper presented the credibility-weighted graph analytics (CWGA) framework for behavioral intention forecasting in disrupted information ecosystems. By introducing credibility propagation as a structured mechanism for maintaining time-varying source reliability scores and coupling these scores to the attention computation of a temporal graph transformer through a disruption-aware gating layer, the framework produces more accurate intention forecasts than methods that treat credibility as static or implicit. Experimental evaluation across three disruption regimes and multiple network scales confirms that the performance advantage is largest under the conditions where accurate forecasting is most consequential: when ecosystem shocks have reorganized the credibility landscape and standard methods produce mis calibrated estimates.

Beyond prediction, the CWGA framework supports interpretable and cost-efficient intervention targeting by identifying nodes that combine structural bridging capacity with current persuasive credibility. The finding that optimal targeting systematically shifts toward institutional sources and cross-community bridges after privacy incidents provides actionable guidance for public health communicators who must respond rapidly to credibility crises with limited resources. The framework is fully reproducible through a controlled simulation environment and provides a principled foundation for future extensions involving real-world data, richer content models, and uncertainty-aware forecasting.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive feedback. M. J. Holloway acknowledges support from the University of North Dakota College of Engineering and Mines research seed grant program. Y. Tan acknowledges support from the Hebei Province Natural Science Foundation (Grant No. F2024201120). B. K. Osei acknowledges support from the University of Cape Coast Research Fund. L. M. Drabik acknowledges support from the Czech Science Foundation project support program.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365
[2] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4-24. https://doi.org/10.1109/TNNLS.2020.2978386
[3] Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224
[4] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001
[5] Khoshraftar, S., & An, A. (2024). A survey on graph representation learning methods. ACM Transactions on Intelligent Systems and Technology, 15(1), Article 11. https://doi.org/10.1145/3633518
[6] Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management Analytics. Nanotechnologies in Construction, 13(3), 181-192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192
[7] Ju, W., Fang, Z., Gu, Y., Liu, Z., Long, Q., Qiao, Z., Qin, Y., Shen, J., Sun, F., Xiao, Z., Yang, J., Yuan, J., Zhao, Y., & Zhu, M. (2024). A

comprehensive survey on deep graph representation learning. Neural Networks, 173, 106207. https://doi.org/10.1016/j.neunet.2024.106207
[8] Barros, C. D. T., Mendonça, M. R. F., Vieira, A. B., & Ziviani, A. (2021). A survey on embedding dynamic graphs. ACM Computing Surveys, 55(1), Article 10. https://doi.org/10.1145/3483595
[9] Lu, Y., Ivanov, L. A., Wang, F., Pisarenko, Z. V., & Ye, C. (2024a). Management analytics: A bibliometric analysis. Nanotechnologies in Construction, 16(3), 257-266. https://doi.org/10.15828/2075-8545-2024-16-3-257-266
[10] Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., King, I., & Pan, S. (2024a). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10466-10485. https://doi.org/10.1109/TPAMI.2024.3443141
[11] Bing, R., Yuan, G., Zhu, M., Meng, F., Ma, H., & Qiao, S. (2023). Heterogeneous graph neural networks analysis: A survey of techniques, evaluations and applications. Artificial Intelligence Review, 56, 8003-8042. https://doi.org/10.1007/s10462-022-10375-2
[12] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. In Proceedings of The Web Conference 2020 (pp. 2704-2710). ACM. https://doi.org/10.1145/3366423.3380027
[13] Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024b). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431-440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440
[14] Xie, Y., Xu, Z., Zhang, J., Wang, Z., & Ji, S. (2023). Self-supervised learning of graph neural networks: A unified review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2), 2412-2429. https://doi.org/10.1109/TPAMI.2022.3170559
[15] Yuan, H., Yu, H., Gui, S., & Ji, S. (2023). Explainability in graph neural networks: A taxonomic survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5782-5799. https://doi.org/10.1109/TPAMI.2022.3204236
[16] Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847
[17] Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., & Wang, S. (2024). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. Machine Intelligence Research, 21(6), 1011-1061. https://doi.org/10.1007/s11633-024-1510-8
[18] Wu, B., Bian, Y., Zhang, H., Li, J., Yu, J., Chen, L., Chen, C., & Huang, J. (2022). Trustworthy graph learning: Reliability, explainability, and privacy protection. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4838-4839). ACM. https://doi.org/10.1145/3534678.3542597
[19] Jin, G., Liang, Y., Fang, Y., Shao, Z., Huang, J., Zhang, J., & Zheng, Y. (2024b). Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. IEEE Transactions on Knowledge and Data Engineering, 36(10), 5388-5408. https://doi.org/10.1109/TKDE.2023.3333824
[20] Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. Expert Systems with Applications, 207, 117921. https://doi.org/10.1016/j.eswa.2022.117921
[21] Jaouadi, M., & Ben Romdhane, L. (2024). A survey on influence maximization models. Expert Systems with Applications, 248, 123429. https://doi.org/10.1016/j.eswa.2024.123429
[22] Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., & Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: A cluster randomised controlled trial. The Lancet, 386(9989), 145-153. https://doi.org/10.1016/S0140-6736(15)60095-2
[23] Li, Y., Gao, H., Gao, Y., Guo, J., & Wu, W. (2023). A survey on influence maximization: From an ML-based combinatorial optimization. ACM Transactions on Knowledge Discovery from Data, 17(9), Article 133. https://doi.org/10.1145/3604559
[24] Centola, D. (2010). The spread of behavior in an online social network experiment. Science, 329(5996), 1194-1197. https://doi.org/10.1126/science.1185231
[25] Kumar, S., Mallik, A., & Panda, B. S. (2022). Influence maximization in social networks using graph embedding and graph neural network. Information Sciences, 607, 1617-1636. https://doi.org/10.1016/j.ins.2022.06.075
[26] Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management Science, 57(9), 1623-1639. https://doi.org/10.1287/mnsc.1110.1421
[27] Hafiene, N., Karoui, W., & Ben Romdhane, L. (2020). Influential nodes detection in dynamic social networks: A survey. Expert Systems with Applications, 159, 113642. https://doi.org/10.1016/j.eswa.2020.113642
[28] Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. Proceedings of the National Academy of Sciences, 109(16), 5962-5966. https://doi.org/10.1073/pnas.1116502109
[29] Kumar, S., Mallik, A., & Panda, B. S. (2023). Influence maximization in social networks using transfer learning via graph-based LSTM. Expert Systems with Applications, 212, 118770. https://doi.org/10.1016/j.eswa.2022.118770
[30] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489, 295-298. https://doi.org/10.1038/nature11421
[31] Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. Science, 334(6060), 1269-1272. https://doi.org/10.1126/science.1207055
[32] Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web (pp. 519-528). ACM. https://doi.org/10.1145/2187836.2187907
[33] Aral, S., & Dhillon, P. S. (2018). Social influence maximization under empirical influence models. Nature Human Behaviour, 2, 375-382. https://doi.org/10.1038/s41562-018-0310-z
[34] Ogburn, E. L., Sofrygin, O., Díaz, I., & van der Laan, M. J. (2020). Causal inference for social network data. Annual Review of Statistics and Its Application, 7, 169-190. https://doi.org/10.1146/annurev-statistics-031219-041813
[35] Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. The Annals of Applied Statistics, 11(4), 1912-1947. https://doi.org/10.1214/16-AOAS1005
[36] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G.,

Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094-1096. https://doi.org/10.1126/science.aao2998
[37] Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. Journalism & Mass Communication Quarterly, 93(1), 59-79. https://doi.org/10.1177/1077699015606057
[38] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559
[39] Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing & Management, 44(4), 1467-1484. https://doi.org/10.1016/j.ipm.2007.10.001
[40] Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences, 116(7), 2521-2526. https://doi.org/10.1073/pnas.1806781116
[41] Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. Journal of the American Society for Information Science and Technology, 53(2), 134-144. https://doi.org/10.1002/asi.10016
[42] Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science, 7(10), 201199. https://doi.org/10.1098/rsos.201199
[43] Westerman, D., Spence, P. R., & Van Der Heide, B. (2014). Social media as information source: Recency of updates and credibility of information. Journal of Computer-Mediated Communication, 19(2), 171-183. https://doi.org/10.1111/jcc4.12041
[44] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological Science, 31(7), 770-780. https://doi.org/10.1177/0956797620939054
[45] Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet information credibility. Journalism & Mass Communication Quarterly, 77(3), 515-540. https://doi.org/10.1177/107769900007700304
[46] Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. Scientific Reports, 10, 16598. https://doi.org/10.1038/s41598-020-73510-5
[47] Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. Journal of Advertising, 19(3), 39-52. https://doi.org/10.1080/00913367.1990.10673191
[48] Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. Nature Reviews Psychology, 1, 13-29. https://doi.org/10.1038/s44159-021-00006-y
[49] Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. Journal of Applied Social Psychology, 34(2), 243-281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x
[50] van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. Nature Medicine, 28, 460-467. https://doi.org/10.1038/s41591-022-01713-6
[51] Sherchan, W., Nepal, S., & Paris, C. (2013). A survey of trust in social networks. ACM Computing Surveys, 45(4), Article 47. https://doi.org/10.1145/2501654.2501661
[52] Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348-384. https://doi.org/10.1080/10463283.2021.1876983
[53] Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems, 43(2), 618-644. https://doi.org/10.1016/j.dss.2005.05.019
[54] Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. Psychological Science in the Public Interest, 21(3), 103-156. https://doi.org/10.1177/1529100620946707
[55] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), Article 109. https://doi.org/10.1145/3395046
[56] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600
[57] Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining fake news: A typology of scholarly definitions. Digital Journalism, 6(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143
[58] Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American Journal of Public Health, 108(10), 1378-1384. https://doi.org/10.2105/AJPH.2018.304567
[59] Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. BMJ Global Health, 5(10), e004206. https://doi.org/10.1136/bmjgh-2020-004206
[60] Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nature Human Behaviour, 5, 337-348. https://doi.org/10.1038/s41562-021-01056-1
[61] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T
[62] Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. Journal of Marketing, 81(1), 36-58. https://doi.org/10.1509/jm.15.0497
[63] Sheeran, P., & Webb, T. L. (2016). The intention-behavior gap. Social and Personality Psychology Compass, 10(9), 503-518. https://doi.org/10.1111/spc3.12265
[64] Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. Journal of Economic Literature, 54(2), 442-492. https://doi.org/10.1257/jel.54.2.442
[65] McEachan, R. R. C., Conner, M., Taylor, N. J., & Lawton, R. J. (2011). Prospective prediction of health-related behaviours with the theory of

planned behaviour: A meta-analysis. Health Psychology Review, 5(2), 97-144. https://doi.org/10.1080/17437199.2010.521684
[66] Baruh, L., Secinti, E., & Cemalcilar, Z. (2017). Online privacy concerns and privacy management: A meta-analytical review. Journal of Communication, 67(1), 26-53. https://doi.org/10.1111/jcom.12276
[67] Sniehotta, F. F., Presseau, J., & Araújo-Soares, V. (2014). Time to retire the theory of planned behaviour. Health Psychology Review, 8(1), 1-7. https://doi.org/10.1080/17437199.2013.869710
[68] Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. Journal of the Academy of Marketing Science, 45, 135-155. https://doi.org/10.1007/s11747-016-0495-4
[69] Betsch, C., Schmid, P., Heinemeier, D., Korn, L., Holtmann, C., & Böhm, R. (2018). Beyond confidence: Development of a measure assessing the 5C psychological antecedents of vaccination. PLOS ONE, 13(12), e0208601. https://doi.org/10.1371/journal.pone.0208601
[70] Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. Information Systems Research, 15(4), 336-355. https://doi.org/10.1287/isre.1040.0032
[71] Brewer, N. T., Chapman, G. B., Rothman, A. J., Leask, J., & Kempe, A. (2018). Increasing vaccination: Putting psychological science into action. Psychological Science in the Public Interest, 18(3), 149-207. https://doi.org/10.1177/1529100618760521
[72] Bélanger, F., & Crossler, R. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. MIS Quarterly, 35(4), 1017-1042. https://doi.org/10.2307/41409971
[73] Larson, H. J., Jarrett, C., Eckersberger, E., Smith, D. M., & Paterson, P. (2014). Understanding vaccine hesitancy around vaccines and vaccination from a global perspective: A systematic review. Vaccine, 32(19), 2150-2159. https://doi.org/10.1016/j.vaccine.2014.01.081
[74] Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C., & Bellekens, X. (2021). Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Computers & Security, 105, 102248. https://doi.org/10.1016/j.cose.2021.102248
[75] MacDonald, N. E. (2015). Vaccine hesitancy: Definition, scope and determinants. Vaccine, 33(34), 4161-4164. https://doi.org/10.1016/j.vaccine.2015.04.036
[76] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010
[77] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
[78] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432
[79] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134. https://doi.org/10.1016/j.knosys.2021.107134
[80] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, 6. https://doi.org/10.1186/s12864-019-6413-7
[81] Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social Networks, 5(2), 109-137. https://doi.org/10.1016/0378-8733(83)90021-7
[82] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243-297. https://doi.org/10.1016/j.inffus.2021.05.008
[83] Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. Physical Review E, 83, 016107. https://doi.org/10.1103/PhysRevE.83.016107