Research on Network Embedding for YouTube Creator Recommendation

Hyeri Lee
Department of Industrial Engineering
Yonsei University
Seoul, South Korea
hyeri0929@yonsei.ac.kr

Seunghwan Song
Department of Industrial & Management Engineering
Korea National University of Transportation
Chungju, South Korea
seunghwan@ut.ac.kr

Haemin Jung*
Department of Industrial & Management Engineering
Korea National University of Transportation
Chungju, South Korea
hmjung@ut.ac.kr

Abstract— **Influencer marketing on YouTube often selects creators by simple popularity indicators (e.g., subscriber count), overlooking whether a creator's core content aligns with a campaign's target topics. We propose a network embedding-based creator recommendation model that integrates social signals—constructed from shared commenters between videos—with textual signals from video titles. The learned representations feed a Content Loyalty index that quantifies a creator's topical adherence for a given keyword. On a dataset of 2,290 videos across 229 keywords with real marketing outcomes, our approach exhibits stronger correlation with Cost per View (CPV) / Cost per Engagement (CPE) than subscriber count and achieves up to 0.829 Area Under to Receiver Operating Characteristic Curve (AUC) on link prediction, indicating high-quality structural representations suitable for recommendation.**

Keywords— Creator Recommendation, Network Embedding, Link Prediction

I. Introduction

YouTube is one of the most dominant social platforms worldwide, where users collectively watch over one billion hours of video every day [1], and surveys show that about 85 percent of U.S. adults use the platform [2]. As the platform continues to expand, influencer marketing has become a crucial strategy for advertisers seeking to enhance brand exposure and consumer engagement. Within this context, the selection of appropriate creators represents a critical decision for advertisers, as it directly determines the alignment between a campaign's message and the audiences most likely to respond to it. The ultimate objective of creator selection is to maximize marketing effectiveness by identifying creators whose content themes and follower bases resonate with the promoted product or service.

However, in practice, this selection process has often been guided by simple popularity indicators—such as subscriber count—rather than by the semantic or thematic relevance of a creator's content to the advertiser's marketing goals. This gap frequently leads to suboptimal campaign outcomes, where highly visible creators fail to deliver meaningful engagement for the intended audience.

To address this issue, this study proposes a network embedding-based creator recommendation model that jointly reflects both the social relationships among videos (via shared commenters) and the textual content of video titles. By integrating these two perspectives, the model learns content-aware video representations that are subsequently aggregated into a Content Loyalty index, enabling advertisers to identify creators whose content is most aligned with their marketing objectives.

Our contributions are threefold:

- We construct a commenter-based video network that captures the latent social proximity among content pieces.
- We design a joint structural–textual network embedding framework that incorporates indirect ties through biased random walks [3].
- We introduce a Content Loyalty index that quantitatively evaluates how well a creator's content aligns with the advertiser's marketing intent, demonstrating higher explanatory power for campaign performance metrics such as Cost per View (CPV) and Cost per Engagement (CPE).

II. Proposed Method

The proposed framework aims to recommend suitable YouTube creators by integrating both the social relationships among videos and their textual content into a unified representation.

*Corresponding author

The overall process consists of three stages: (1) constructing a video-level network, (2) learning content-aware embeddings through joint structural–textual modeling, and (3) computing a Content Loyalty index for creator recommendation.

A. Video Network Construction

We first construct a video-level graph $G = (V, E, T)$, where each node $u \in V$ represents a video, each edge $(u, v) \in E$ indicates that the two videos share at least one common commenter, and T stores the textual metadata such as video titles. This graph captures the implicit viewer overlap among creators and serves as the foundation for embedding learning. The objective is to learn a low-dimensional vector z_u for each node that preserves both structural proximity and textual similarity within the network.

B. Text Encoder

For each video title $S_u = (w_1, \dots, w_n)$, we employ an embedding lookup followed by a bidirectional Long Short-Term Memory (LSTM) [4] with a word-level attention mechanism. This process captures contextual dependencies between words while emphasizing informative terms using attention weights computed through a global context vector and SoftMax normalization. The resulting vector t_u represents the text-based embedding of node u.

C. Structure Encoder with Indirect Ties

While direct edges encode explicit relationships, important creator associations may exist through indirect connections. To capture both, we apply node2vec-style biased random walks [3] that generate node sequences under parameters p and q, balancing breadth- and depth-first exploration. From these sequences, a structure-based embedding s_u is learned to predict neighboring nodes. We jointly optimize four directional likelihoods— $s \rightarrow s$, $t \rightarrow t$, $s \rightarrow t$, and $t \rightarrow s$ —following context-aware designs [5], [6], aligning structural and textual spaces for more coherent representations.

D. Joint Objective and Optimization

Given $u \in \{s_u, t_u^{(v)}\}$ and $v \in \{s_v, t_v^{(u)}\}$, the objective maximizes

$$\log \sigma(u^\top v) + \sum_{i=1}^{k} \mathbb{E}_{z \sim P(v)}[\log \sigma(-u^\top z)] \qquad (1)$$

where $P(v) \propto \text{degree}(v)^{3/4}$ denotes the negative sampling distribution [7]. The entire model is optimized using the Adam algorithm [8]. This joint objective encourages embeddings to capture both network structure and content-level semantics, enabling more accurate creator matching.

E. Content Loyalty for Creator Recommendation

Once embeddings are obtained, we quantify how closely each creator's content aligns with a given campaign keyword. For a keyword K, we retrieve its top-N search videos and compute the cosine similarity between their embeddings and those of the creator's videos. Let p denote a similarity threshold. The Content Loyalty index is defined as

$$L_m(C, K, p) = \frac{n_{\geq p}}{N} \bar{s} \qquad (2)$$

where $n_{\geq p}$ denotes the number of videos produced by creator C whose similarity to videos associated with keyword Kexceeds the threshold p; N is the total number of videos produced by creator C; m indicates the aggregation method (average or top-2); and $\bar{s}$ represents the aggregated similarity score. This formulation provides a balance between coverage (how many videos are relevant) and intensity (how strongly they match the campaign theme). Creators are ranked according to their $L_m(C, K, p)$ values, allowing advertisers to identify those with the highest thematic consistency and potential marketing effectiveness.

III. EXPERIMENTS

To evaluate the effectiveness of the proposed framework, we conducted experiments using real YouTube data and campaign performance records. The evaluation focuses on two aspects: (1) how well the proposed Content Loyalty index explains marketing outcomes such as CPV and CPE; and (2) the quality of learned embeddings, measured through a link prediction [9] task.

A. Dataset

We collected data through the YouTube API by querying 229 distinct marketing keywords and retrieving the Top-10 videos per keyword, yielding 2,290 videos in total. For each video, we gathered metadata including the channel ID, title, like count, and commenter IDs to construct the commenter-based network described in Section II. Additionally, we obtained real campaign logs from advertiser partnerships, containing CPV and CPE values over a six-week period. Only creators whose campaigns had measurable performance data were included in the evaluation. This dataset enables an integrated analysis of both social relationships (based on shared commenters) and economic performance metrics for recommendation validation.

B. Evaluation Metrics and Settings

Two quantitative evaluations were performed:

- **Marketing correlation analysis.** We computed Pearson correlations between each creator-level predictor and the observed CPV/CPE values. The baseline predictor was the creator's subscriber count, while the proposed predictor was the log-transformed Content Loyalty index under various parameter settings: similarity threshold $p \in \{0.4, 0.5, 0.6, 0.7\}$ and aggregation method $m \in \{avg, top2\}$.
- **Link prediction.** To examine the representational quality of the learned embeddings, we performed a standard link-prediction task. A portion of the edges (15 % to 95 %) was randomly removed from the graph and used as the test set, while the remaining edges served as training data. Performance was measured by the Area Under the ROC Curve (AUC), where higher values indicate better structural preservation.

C. *Results and Discussion*

Table 1 summarizes the correlation between marketing outcomes and different predictors. Under higher values of p, the Content Loyalty index exhibits stronger negative correlations with both CPV and CPE than subscriber count, indicating that creators with higher loyalty scores tend to achieve more cost-efficient engagement. The strongest associations are observed when $p \approx 0.6$–0.7 and the $top2$ aggregation is applied, yielding correlations of approximately -0.42 for CPV and -0.36 for CPE, whereas the subscriber-based baseline shows more moderate correlations of -0.28 and -0.30, respectively. These results suggest that the proposed embedding-based loyalty measure captures aspects of semantic content alignment that are more closely related to advertising efficiency than popularity-oriented indicators.

TABLE I. CORRELATION BETWEEN PREDICTORS AND CAMPAIGN OUTCOMES (CPV, CPE)

	CPV	**CPE**
subscriberCount	-0.283	-0.304
log $L_{avg}(0.4)$	-0.232	-0.232
log $L_{avg}(0.5)$	-0.215	-0.206
log $L_{avg}(0.6)$	**-0.340**	-0.285
log $L_{avg}(0.7)$	**-0.340**	-0.294
log $L_{top2}(0.4)$	-0.288	-0.274
log $L_{top2}(0.5)$	-0.269	-0.257
log $L_{top2}(0.6)$	**-0.419**	**-0.359**
log $L_{top2}(0.7)$	**-0.391**	**-0.332**

For the link prediction task, Table 2 reports the Area Under to Receiver Operating Characteristic Curve (AUC) results under varying proportions of training edges. The proposed model achieves the best performance among all baselines across all edge ratios except for the 15% setting. This steady improvement demonstrates that incorporating both textual information and indirect structural links effectively enhances network representation learning.

TABLE II. LINK PREDICTION PERFORMANCE

Training Edge Ratio	**CANE (Text Only)**	**CANE**	**CANE + Random Walk**	**Proposed Model**
0.15	0.593	0.650	**0.736**	0.730
0.35	0.677	0.748	0.771	**0.780**
0.55	0.706	0.760	0.788	**0.808**
0.75	0.710	0.769	0.812	**0.818**
0.95	0.729	0.789	0.820	**0.829**

IV. CONCLUSION

This study proposed a network embedding–based framework for YouTube creator recommendation that integrates social interaction signals and textual semantics to better capture content relevance in advertising contexts. By learning unified content-aware embeddings from commenter-based video networks and title-level textual representations, we introduced the Content Loyalty index as a quantitative measure of how consistently a creator's content aligns with a given marketing keyword. Empirical evaluations demonstrated that this loyalty-based measure exhibits a stronger association with campaign performance metrics, such as CPV and CPE, than conventional popularity indicators, while the embedding model itself showed strong representational quality in link-prediction tasks. These results highlight the effectiveness of embedding-driven representations in modeling both semantic alignment and relational structure among creators.

From a practical perspective, the proposed framework provides advertisers with a principled, data-driven approach to creator selection that goes beyond surface-level popularity and focuses on content relevance and consistency. By grounding creator recommendation in learned representations and loyalty-oriented metrics, the approach enables more meaningful alignment between marketing objectives and creator content. Future work will extend this framework by incorporating richer multimodal signals and by generalizing the model to heterogeneous creator–content–viewer graphs using advanced representation learning paradigms, further enhancing its applicability to real-world advertising ecosystems.

REFERENCES

[1] BroadcastPro, "YouTube touches one billion daily hours of video consumption," Dec. 2024. Available: https://www.broadcastprome.com/news/youtube-touches-one-billion-daily-hours-of-video-consumption/.

[2] Pew Research Center, "5 facts about Americans and YouTube," Feb. 2025. Available: https://www.pewresearch.org/short-reads/2025/02/28/5-facts-about-americans-and-youtube/.

[3] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016.

[4] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, 1997.

[5] C. Tu, H. Liu, Z. Liu, and M. Sun, "CANE: Context-aware network embedding," Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2017.

[6] J. Liu, H. Liu, Y. Zhou, and Z. Liu, "Hierarchical Attention-Based Semi-supervised Network Representation Learning," Proc. NLPCC, 2018.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2015.

[9] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," J. Amer. Soc. Inf. Sci. Technol., vol. 58, no. 7, pp. 1019–1031, 2007.