Integrating Sequential and Relational Modeling for User Events: Datasets and Prediction Tasks

(/pdf?id=9rCQVo4KXg)

07 May 2025 (modified: 18 Sept 2025) Submitted to NeurIPS 2025 Datasets and Benchmarks Track Datasets and Benchmarks Track, Senior Area Chairs, Area Chairs, Reviewers, Authors

Keywords: User Events, Graph, Sequence, GNN, Transformers, Datasets

Abstract:
User event modeling plays a central role in many machine learning applications, with use cases spanning e-commerce, social media, finance, cybersecurity, and other domains. User events can be broadly categorized into personal events, which involve individual actions, and relational events, which involve interactions between two users. These two types of events are typically modeled separately, using sequence-based methods for personal events and graph-based methods for relational events. Despite the need to capture both event types in real-world systems, prior work has rarely considered them together. This is often due to the convenient simplification that user behavior can be adequately represented by a single formalization, either as a sequence or a graph. To address this gap, there is a need for public datasets and prediction tasks that explicitly incorporate both personal and relational events. In this work, we introduce a collection of such datasets, propose a unified formalization, and empirically show that models benefit from incorporating both event types. Our results also indicate that current methods leave a notable room for improvements. We release these resources to support further research in unified user event modeling and encourage progress in this direction.

Dataset Submission: This submission includes a dataset.
Croissant File: json (/attachment?id=9rCQVo4KXg&name=croissant_file)
Dataset URL: https://kaggle.com/datasets/8c7dcbcc4b8aa0ed72d4db8589edb2a1cae2e792bcefa359190ca4e2d133c12e (https://kaggle.com/datasets/8c7dcbcc4b8aa0ed72d4db8589edb2a1cae2e792bcefa359190ca4e2d133c12e)
Dataset Final Confirmation: true
Checklist Confirmation: I confirm that I have included a paper checklist in the paper PDF.
Code URL: https://anonymous.4open.science/r/4bade975a17348b39d07d6b1d0ece30b/ (https://anonymous.4open.science/r/4bade975a17348b39d07d6b1d0ece30b/)
Responsible Reviewing: We acknowledge the responsible reviewing obligations as authors.
Primary Area: Datasets & Benchmarks illustrating Different Deep learning Scenarios (e.g., architectures, generative models, optimization for deep networks, foundation models, LLMs)
LLM Usage: Editing (e.g., grammar, spelling, word choice), Implementing standard methods, Other (enter in the text field below)
Other LLM Usage: We use LLM mainly for writing assistance (such as revising our writing and constructing a paragraph draft based on the key talking points we provided to the LLM) and coding assistance (such as, debugging, suggesting code based on our specification, and helping us implementing the baseline methods).
Declaration: I confirm that the above information is accurate.
Submission Number: 705

Discussion (?id=9rCQVo4KXg#discussion)

Filter by reply type... Filter by author... Search keywords... Sort: Newest First

Everyone Program Chairs Submission705 Authors Submission705... Submission705 Area... Submission705... Submission705... Submission705... Submission705...

17 / 17 replies shown

Add: **Withdrawal**

Paper Decision

Decision by Program Chairs 18 Sept 2025, 08:23 (modified: 18 Sept 2025, 12:58) Program Chairs, Authors

Revisions (/revisions?id=TZlen6Aa4T)

Decision: Reject

Comment:

This paper introduces a framework, Personal and Relational Event Sequence (PRES), to model user events by integrating personal and relational interactions, along with curated large-scale datasets. While the unified framework and benchmark tasks are interesting and address a gap in event modeling, the work has notable limitations. First, the datasets are derived from existing public data rather than newly collected or annotated resources, limiting the originality of the contribution. Also, the pipeline for dataset processing is unclear, and accessibility issues with dataset files undermine reproducibility. The lack of field-level descriptions and detailed comparisons with related datasets, such as Amazon Reviews or Reddit, further weakens the clarity and distinction of the contribution. The methodology appears to unnecessarily repackage existing tasks like dynamic graph modeling without adding substantial utility. The benchmarks rely on relatively outdated methods, and the absence of multimodal tasks reduces relevance to modern research.

Official Review of Submission705 by Reviewer 55RT

Official Review by Reviewer 55RT 02 Jul 2025, 14:25 (modified: 24 Jul 2025, 09:04)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 55RT

Revisions (/revisions?id=YX01xhsP84)

Summary:

The submission introduces a unified framework for modeling user events that integrates both personal events (individual actions) and relational events (user-to-user interactions). The authors propose the Personal and Relational Event Sequence (PRES) formalization to jointly capture these event types while preserving temporal structure. They curate and release several large-scale datasets from domains like e-commerce, social media, and software development, along with benchmark tasks for predicting both personal and relational events. Experiments show that combining both event types improves predictive performance, while existing models struggle in this unified setting. The work provides new resources to advance research on holistic user event modeling.

Strengths Contributions:

The submission's main strength is its novel integration of personal and relational user events through the Personal and Relational Event Sequence (PRES) formalization, addressing a clear gap in existing datasets and benchmarks. Unlike prior resources such as TGB or OGB, which focus on relational data, or sequential datasets that ignore user interactions, this work provides large-scale, diverse datasets that capture both event types (e.g., pres-amazon, pres-brightkite). The authors clearly justify their distinction from prior work and demonstrate through experiments that combining both event types improves performance. The paper is well-organized and clearly written, with informative tables (e.g., Table 1) and figures (e.g., Figure 1) that effectively convey the core ideas. This resource has strong potential to advance research on unified user event modeling.

Limitations Weaknesses:

The datasets, while diverse across domains, are built from existing public data reformatted for PRES (Section 4.1), which may not fully capture the complexity of real-world user behavior. Future work could create datasets with richer attributes and event semantics. The formalization lacks support for event- or user-level features (Section 6), limiting its applicability for feature-rich modeling tasks; extending the framework to include such features would broaden its utility. The scalability of models is also a concern: temporal graph methods struggled with large datasets like pres-github (Table 3), suggesting a need for more efficient or hybrid approaches. Lastly, while the paper is well-organized, dense tables (e.g., Tables 2–4) could benefit from clearer visual summaries to highlight key findings.

Ethical Considerations: No, there are no or only very minor ethics concerns

Dataset Code Accessibility: Yes

Rating: 5: Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate-to-high impact on more than one area of AI, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct Acknowledgement: Yes

Responsible Reviewing Acknowledgement: Yes

Rebuttal by Authors

Rebuttal by Authors 31 Jul 2025, 03:02 (modified: 31 Jul 2025, 18:55)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=IZV1wDTZGt)

Rebuttal:
We thank the reviewer for their positive feedback and useful comments. We appreciate that they found our work to be technically solid, and recognized its potential impact. We will incorporate the following changes in our revision:

Dataset Curation

We agree. Our primary goal was to create the first benchmark that unifies the two event types from available data to establish a foundation for this research direction. PRES was explicitly designed to be a template: the released preprocessing code makes it easier for others to create PRES from their richer event and user semantics. We view this as a valuable direction for future work.

Richer attributes and event semantics

Our current PRES setup was intentionally kept simple to establish a baseline. We agree that extending it to support richer features is an important next step. For example, a personal event like a product rating could be included with features like the product category/price, full review (if provided), and others. A relational event like a social network friendship could include features like shared interest, friendship duration, or friendship discovery status. Finally, we could also include user level features such as user demographic or preference. We will add a discussion of these extensions to the paper.

Scalability of temporal graph methods

We agree that temporal graph methods (TGN, DyRep) struggle with large-scale datasets such as pres-github. Improving scalability is an open challenge in temporal modeling. We can add a short discussion on potential directions as future work: (i) hybrid approaches that combine sequential modeling (e.g., Transformer-based encoders) with sampled temporal neighborhoods, (ii) graph sparsification and pruning of less informative edges/events, (iii) mini-batching strategies and approximate memory mechanisms for TGN-style models.

Dense tables (Tables 2–4)

We appreciate this suggestion and agree that our tables in the main paper can be simplified. Additionally, we do include some visualizations in Appendix C (Figures 4 & 5), but these could be made more prominent.

For the final version, we plan to: (i) reduce metrics to MRR and hits@k for fewer numbers of k, (ii) add delta MRR (relative improvement over baseline) as an additional column, (iii) move the original dense tables to the appendix along with referring to Figures 4 and 5 (bar plots) as an alternative visual overview of the results.

As an example, improved Table 2 for relational prediction tasks on pres-brightkite, the GAT+S model will show a +23.8% delta MRR improvement over the GAT baseline. Improved Table 4 for personal prediction on pres-gowalla, the TGN-PR model (using both event types) will show a +30.8% delta MRR compared to the TGN model (using only personal events), demonstrating the benefit of unified approaches.

Mandatory Acknowledgement by Reviewer 55RT

Mandatory Acknowledgement by Reviewer 55RT 04 Aug 2025, 03:56

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Mandatory Acknowledgement by Reviewer 55RT

Official Comment by Reviewer 55RT

Official Comment by Reviewer 55RT 08 Aug 2025, 19:30

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thank you for the clarification. I am satisfied with your response and will maintain my positive score.

Official Review of Submission705 by Reviewer 8ppy

Official Review by Reviewer 8ppy 01 Jul 2025, 11:20 (modified: 18 Sept 2025, 13:03)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 8ppy

Revisions (/revisions?id=JoEwJDo17w)

Summary:
This work introduces a unified approach to modeling both personal and relational user behaviors, which are two aspects often studied separately in the literature. The authors introduce five real-world datasets spanning multiple domains, each supporting a range of standardized prediction tasks. Empirical results demonstrate that jointly modeling both event types improves performance, underscoring the importance of integrated approaches.

Strengths Contributions:
1.The paper is clearly written and well-structured. Both the code and datasets are publicly available.

2.The authors introduce five diverse, real-world datasets covering multiple domains, offering a valuable and reusable benchmark suite for future research.

3.The dataset integrates both personal (sequential) and relational (graph-based) user events, enabling unified modeling of heterogeneous user behaviors. This design reflects real-world interaction complexity and is relatively valuable among existing benchmarks.

4.Experimental results demonstrate that jointly modeling both event types improves performance across tasks, highlighting the benefit of integrated approaches.

Limitations Weaknesses:
1.The dataset lacks field-level description, with 0 out of 345 fields annotated according to the metadata report. This limits clarity and usability for researchers seeking to reuse or extend the benchmark.

2.A substantial portion of the dataset files (51 in total) are reported as inaccessible or broken, undermining claims of reproducibility and potentially limiting adoption by the community.

3.The paper does not provide a thorough comparison with related datasets such as Amazon Reviews or Reddit datasets. A clearer discussion of how this benchmark differs in scope, scale, or task design would help establish its contribution relative to prior resources.

4.The datasets appear to be compiled from existing public sources. While the unification of personal and relational events under a consistent schema is valuable, the absence of newly collected or annotated data somewhat limits the originality of the contribution from a dataset creation standpoint.

5.The paper would benefit from a deeper analysis of the observed patterns. For example, it remains unclear why most methods perform better under the "+S" configuration or why the removal of "-RP" tends to hurt performance.

6.Minor: All benchmark tasks focus on structured event data. As many modern systems and models incorporate multimodal inputs (e.g., text, images), expanding the benchmark to support such modalities could broaden its impact and relevance.

Ethical Considerations: No, there are no or only very minor ethics concerns
Dataset Code Accessibility: Partly
Dataset Code Comments:
While the submission provides both code and datasets, there are several issues that limit full accessibility and reproducibility:

1.According to the Croissant metadata report, none of the 345 dataset fields include descriptions.

2.A total of 51 files in the dataset repository are reported as inaccessible or broken. This raises serious concerns about reproducibility.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:
Rebuttal addresses most of my concerns, including thorough comparisons with related datasets, deeper analysis of the observed patterns. As a result, I maintain my positive rating.

Rebuttal by Authors

Rebuttal by Authors 31 Jul 2025, 01:20 (modified: 31 Jul 2025, 18:55)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=qzauWoRx7M)

Rebuttal:

We sincerely thank the reviewer for the thoughtful and positive review. We appreciate the recognition of our paper's clarity, the value of the curated datasets, and the importance of jointly modeling personal and relational events. We also appreciate the thorough critical feedback provided, which we would like to address as follows.

Missing field level description

We thank the reviewers for pointing out the missing field-level description in our dataset's metadata. Our datasets are stored in csv files with this column format:

- uid: user ID
- timestamp: event timestamp
- event_set: event set. either "personal" or "relational"
- event: event name
- other_uid: another user ID (for relational events)

All the csv files follow the same conventions. The many missing field level descriptions come from the fact that each dataset is stored in multiple files, e.g. one file for each train set, val set, test set for personal/relational events, negative sample set; for each dataset and each prediction task.

We have provided the field description above in our dataset's readme file. However, as correctly pointed out by the reviewer, we did not add the description in the metadata of each individual file when we uploaded our dataset in the Kaggle website. We will add these metadata in the revised version of our paper/dataset.

Inaccessible files

During this review process, we chose a double-blind review process, instead of the single-blind process. For that, we decided to not release the dataset into the public yet during the review process. We selected sharing the dataset via Kaggle's private URL preview link sharing following the D&B track FAQ.

> **I don't want to make my dataset publicly accessible at the time of submission. What are my options?**
> Harvard Dataverse and Kaggle platforms both offer private URL preview link sharing. This means your dataset is only accessible to those with whom you share its special URL, e.g., reviewers. Note that you will be required to make your dataset public by the camera-ready deadline. Failure to do so may result in removal from the conference and proceedings.

This likely caused the reported inaccessible files during the Croissant file validation, since it may not work with private URL preview link sharing. In the final version of our paper, we will release the dataset publicly, and make sure that all the files (including metadata) are accessible by the public.

Comparison with other datasets

We thank the reviewer for raising this point. We will add a more detailed comparison of our dataset versus previously published datasets. Below is an example on the comparison with the datasets contributed by the TGB paper:

TGB (Huang, et.al, 2023) contains several datasets that model user behaviors. These datasets can be roughly divided into two categories: (1) user-to-user interaction datasets, such as tgbl-coin, tgbl-comment and tgbn-trade; (2) user-to-item bipartite interaction datasets, such as tgbl-wiki, tgbl-review, tgbn-genre, tgbn-reddit, and tgbn-token. The first category of user-to-user interaction datasets are similar to the relational event part of our datasets; however our datasets also contain personal event sequences, in addition to relational events, which are not present in the TGB dataset.

The second category of TGB datasets are bipartite temporal graphs. In our PRES formulation, the interaction of a user to an item can be encoded as a personal event, where an item is represented as an event or a token. However, the personal event abstraction in the PRES formulation can also encode other types of events. An example is illustrated in Figure 1, where User A has both user-item events (product views) and other types of events such as 'Add to cart' and 'Purchase'. Our formulation also contains relational events that model use-to-user interactions.

In addition, formulating an item as a personal event instead of a node enables the flexibility of encoding items that have hierarchical information such as Geohash in our Brightkite dataset. Each character in the geohash encodes increasingly detailed location information. When we encode an 8-letter geohash as a node, we lose the hierarchical information encoded in the geohash. In contrast, if we are not forced to represent an event as a node, we have more flexibility to encode the hierarchical structure of the geohash. For example, in a sequence model, one could tokenize the event freely. A single event could be encoded into multiple tokens.

In terms of the size, the TGB dataset ranges from a small size of 255 node graph (tgbn-trade) to nearly a million node graph (tgbl-comment). Our PRES dataset also ranges from a small-to-medium size of 58 thousand users (pres-brightkite) to a relatively large dataset of pres-github with 3.6 million users. In terms of the number of events (or number of edges in TGB dataset) our PRES datasets are comparable with TGB datasets, and in some cases larger than TGB datasets. The number of edges in TGB datasets range from around 150 thousands edges (tgbl-wiki) to 44 million edges (tgbl-comment); whereas the number of events in our datasets range from 1.5 million (pres-amazon-clothing) to more than 100 million events (pres-github).

We will add more comparisons with other benchmark datasets (such as TGB 2.0 (Gastinger, et.al, 2024)) in our revised paper.

Data curation instead of data creation

We agree with the reviewer's assessment that the contribution of the paper lies in curating user event datasets with the property of containing both personal and relational events, as well as in preparing prediction tasks for those datasets. We do not create entirely new datasets. Instead, we process and reformulate raw datasets into PRES datasets and corresponding prediction tasks. However, we would like to note that many prior works in dataset and benchmark tracks, such as OGB (Hu, et.al, 2020), TGB (Huang, et.al, 2023), and TGB 2.0 (Gastinger, et.al, 2024), also focus on dataset curation rather than creation. We think that this should not diminish the contribution of our paper within the dataset and benchmark track.

Deeper analysis of the results

The main takeaway of the paper is that the best models for both personal and relational events outperform those using only one in either relational or personal event task predictions. For example in relational event prediction +S models (GCN+S and GAT+S) add the sequence embedding of personal events data into the relational event graph; boosting the performance over the models that use relational events only (GCN and GAT). This shows the need of building models that take into account both signals.

We would like to also clarify that GCN-RP and GAT-RP models do not drop any information, the difference between GCN and GCN-RP is that GCN only operates on relational events data; whereas in GCN-RP, we convert each unique personal event into a node and add it to the user graph used by GCN model, creating edges between users and their personal event nodes. In some datasets, like Brightkite, Gowalla, and Github, adding personal event nodes to the graph reduces their performance. This could be explained as follows:

- When we exclude personal events (GCN and GAT), the model is able to extract and learn some predictive information from the relational events alone to some degree.
- When we include personal events as nodes (GCN-PR, and GAT-PR), this ends up adding more noise than benefit into the system; as the number of personal event nodes is much larger than the number of user nodes. The result is that the model performance decreases.
- When we encode personal events as a sequence embedding (GCN+S and GAT+S), this creates meaningful features without introducing excessive noise. The models are able to pick up on additional signals from these personal event embeddings, and performance increases.

This pattern does not generalize to all datasets, as we can see from Amazon-clothing and Amazon-electronics datasets where they perform relatively well on MRR and Hits@5, but not on Hits@k metrics with larger k. This may show that on these two datasets, the personal event nodes do not merely just become noise in the graphs. Instead, it helps in getting the model to improve at the precision on the top candidate (i.e. fewer but more accurate suggestions), with the expense of getting lower recall coverage on the larger k. This is an interesting observation that may influence the architecture design of future models that want to take advantage of both personal and relational event signals.

We will add additional similar analysis of our results in our revised paper, either in the main paper or in the appendix section.

Multimodel expansion of the datasets

This is a great point. Yes, the focus of this paper is on user events on structured data. Adding multimodal events such as text and events could be an interesting future work to explore.

Mandatory Acknowledgement by Reviewer 8ppy

Mandatory Acknowledgement by Reviewer 8ppy 03 Aug 2025, 16:26

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I

understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Official Comment by Reviewer 8ppy

Official Comment by Reviewer 8ppy 03 Aug 2025, 16:29

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thanks for the rebuttal. I decide to maintain my positive score.

Replying to Official Comment by Reviewer 8ppy

Official Comment by Authors

Official Comment by Authors 03 Aug 2025, 19:31

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
We thank the reviewer for maintaining the positive recommendation to our paper. We would be happy to provide any additional information about the detail of our paper, if needed. Thank you.

Official Review of Submission705 by Reviewer wgDA

Official Review by Reviewer wgDA 10 Jun 2025, 12:32 (modified: 18 Sept 2025, 13:03)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer wgDA

Revisions (/revisions?id=V5J1kpfUaz)

Summary:
The paper proposes a unified framework for user event modelling. The paper defines prediction tasks for both relational and personal event forecasting, and evaluate a range of baseline models—including graph-based and sequence-based. The paper also releases five benchmark datasets for the formalised problem.

Strengths Contributions:

1. The paper proposes a new problem or task for user events data.
2. The paper is well-organised and easy to understand.
3. The paper discusses several methods which could be applied to the proposed problem.

Limitations Weaknesses:

1. The pipeline of the dataset processing is not clearly described or not mentioned.
2. The benchmark methods used in the experiments are relatively outdated, and more recent state-of-the-art models could have been included for comparison.
3. From my perspective, the problem formalisation is not quite necessary. The tasks defined on these user event datasets can easily be addressed using existing dynamic graph models or time series forecasting models. The proposed formalisation seems to unnecessarily complicate and repackage previous tasks such as node classification or link prediction.
4. The proposed datasets are derived from existing public datasets rather than newly collected ones.

Ethical Considerations: No, there are no or only very minor ethics concerns
Dataset Code Accessibility: Yes
Dataset Code Comments:
The datasets are available on Kaggle and code is available on GitHub repo.

Rating: 3: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:

Based on my overall assessment of the paper and the rebuttal, I have decided to raise my score to borderline reject.

Rebuttal by Authors

Rebuttal by Authors 31 Jul 2025, 03:01 (modified: 31 Jul 2025, 18:55)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=pCKF6bb6vA)

Rebuttal:

We thank the reviewer for their thorough reviews, comments, and concerns. We appreciate the reviewer's assessment that our paper is well organized and easy to understand. We will address the concerns raised as follows.

Problem formalisation

We respectfully disagree with the reviewer's assessment that the prediction tasks discussed in our paper can be easily addressed using existing dynamic graph models or time series forecasting models. In each of our prediction tasks, we utilize both personal and relational events as predictors. In contrast, while dynamic graph models are a great fit for modeling relational events, they do not provide effective tools for modeling personal events. On the other hand, sequence or time series forecasting models are able to pick up signals from personal event sequences but lack any means to integrate interactions with other users via relational events.

For illustraton, we highlight an example of user activities on an e-commerce website in Figure 1. The relational events between User A and User B: User A sending gift money to User B. All other events from User A are personal events. Let's see how existing models or frameworks would handle this setup.

1. Dynamic graph model

Many dynamic graph models, like CTDG-based models, define the prediction task as a stream of triplets (u_i, v_i, t_i), where each triplet is an interaction between users u_i and v_i at time t_i. This works well for relational events, like User A sending gift money to User B. However, it cannot easily capture personal events, since everything must be represented as nodes, and events are treated only as interactions between two nodes.

Dynamic graph models would have difficulty representing events such as "Login via App," "Search," "Write Review," or other possible events like "Change Password," "Change Username," "Add New Address," "Join Subscription," "Make Subscription Payment," "Payment Successful," or "Payment Declined."

Of course, there is a workaround: converting personal events into nodes and then building a heterogeneous dynamic graph representation that encodes the event as an interaction between a user node and this newly created event node. In some cases, like viewing products, this approach may make sense; so the personal event is converted into an edge between User A and Product B. However, this workaround will not work for all types of events. Events like "Login via App," "Make subscription payment," "Payment successful," or "Payment declined" are not naturally represented as nodes. These events reflect a user's status change or actions that don't involve others and because there's no second entity, modeling them as graph edges feels unnatural.

2. Sequence or time series model.

Sequence-based models can naturally handle personal event sequences, but they are not good at representing relational events. One option is to flatten the relational information, for example, turning "User A sends gift money to User B" into just "Sending gift money." But this removes information about which users interacted with User A. For some prediction tasks, it's important to include the behavior of other users who interacted with User A, which this approach cannot capture.

Other issues

There are also other issues that could arise when we model user events using existing models/formalizations. Among them are:

- In graph models, every entity must be a node. This becomes a problem when the entity has a hierarchical structure. For example, in the Brightkite dataset, users' personal events are check-ins at different locations, encoded using Geohash level 8. Each character in a geohash adds more detail about the location. The first letter gives a rough area (±2500 km), four letters give about ±20 km, and eight letters narrow it down to about ±19 meters. Geohashes with more shared prefix characters are geographically closer than those with fewer.

  When we turn an 8-letter geohash into a node in a graph model and use it in a dynamic graph, we lose the hierarchical information in the geohash. In contrast, if we don't have to represent events as nodes, we gain more flexibility. For example, in a sequence model, we can freely tokenize events. A check-in event can be split into multiple tokens, like breaking the geohash into four parts, each with 2 characters. This way, we can still keep the hierarchical structure of the geohash.
- When building models for user prediction tasks, it helps to use both personal and relational events. For example, in Brightkite, to recommend a new friend (a relational task), it's useful to know the user's check-in history (personal events). And to recommend the next check-in location (a personal task),

knowing the user's friendships (relational events) is also helpful. Our experiments support this: models that use both types of events usually perform the best.

Data processing

Thank you for the helpful feedback on the data processing stage. While we include some details in Section 4, we agree that the explanation is not clear enough. In our revised paper, we will add a more detailed description of the data processing steps in the appendix.

In short, the data processing files are available in our codebase. Each dataset has a create_datasets/process_X.ipynb file for generating processed data, and a create_datasets/task_X.py file for creating prediction tasks, where X is the dataset name. Due to space limits, we will explain only the processing for pres-brightkite in this rebuttal.

The main script is process_X.ipynb, which formats the data so that each row is a personal or relational event. For pres-brightkite, the steps are:

- The raw data of users' friendships and check-in histories are stored in text files. We read the friendship files into a dataframe and reformat them into our standard format.
- The check-in history file contains the latitude and longitude of check-in locations. After reading the files, we convert the latitude-longitude encoding of locations into Geohash-8 string representations and convert the check-in times into Unix timestamp format
- We then combine this with the relational events, sort the data by uid and timestamp, and save it into a CSV file (in our dataset repository on Kaggle, this is processed/brightkite_all_events.csv)

Once all events are saved in a standard format, we use create_datasets/task_brightkite.py to create data splits and pre-generate negative samples for reproducibility. For both tasks, we first split the events into two dataframes: one for personal events and one for relational events.

To create the relational event prediction task, we randomly split the relational event dataframe into 70% train, 10% eval, and 20% test sets, and save them as CSV files. Negative samples are generated by randomly picking users, making sure to exclude those who already have relational events with the target user in the training set.

To generate the personal event prediction task, we perform timestamp-based splitting on the personal event dataframe. We split each user's personal events by taking the last 20% for the test set, the previous 10% for validation, and the remaining 70% for training. We cap the number of events in the test and validation sets to at most 20 and 10 per user, respectively. Negative samples are generated by stratified hierarchical sampling across varying extents of geohash character matching, ensuring that our negatives contain a mix of nearby and distant locations.

More recent baselines

We chose the baseline models based on widely used architectures: GCN and GAT for static graphs, TGN and DyRep for temporal graphs (based on TGB (Huang et al., 2023) and TGB 2.0 (Gastinger et al., 2024)), and BERT for sequence models. Based on the reviewer's feedback, we added more recent baselines: GATv2 (Brody et al., 2021) and TransformerConv (Shi et al., 2020) for static graphs, and TNCN (Zhang et al., 2024) for temporal graphs. Below are the experimental results for the relation prediction task. All results are averaged over 5 runs, except TNCN, which is averaged over 3 runs.

Dataset	Model	MRR (%)	Hits@10 (%)
Brightkite	GATv2	32.9	56.3
Brightkite	GATv2+S	41.2	65.4
Brightkite	GATv2-RP	11.3	22.0
Brightkite	TransformerConv	40.2	63.5
Brightkite	TransformerConv+S	47.4	71.4
Brightkite	TransformerConv-RP	15.8	29.2
Brightkite	TNCN	28.2	40.3
Gowalla	GATv2	39.2	65.0
Gowalla	GATv2+S	42.6	68.6
Gowalla	GATv2-RP	14.0	26.7
Gowalla	TransformerConv	47.4	72.1
Gowalla	TransformerConv+S	49.9	74.4
Gowalla	TransformerConv-RP	21.0	38.1
Gowalla	TNCN	25.8	34.9

The results support our main findings. Among static graph models, TransformerConv performs better than GCN and GAT on both datasets. Adding sequence embeddings (TransformerConv+S) further improves performance, similar to what we saw with GCN+S and GAT+S. However, adding personal event nodes to the graph (TransformerConv-RP) lowers performance, as also seen in GCN-RP and GAT-RP. For temporal graph models, TNCN clearly outperforms TGN and DyRep on both datasets. But because TNCN uses a suboptimal graph structure that includes personal event nodes in the relational graph, it still performs worse than the "+S" models.

We will continue running these additional experiments on other datasets and include the results in our revised paper.

Data curation instead of data creation

We agree with the reviewer that our main contribution is curating user event datasets that include both personal and relational events, and creating prediction tasks for them. We do not build entirely new datasets, but instead process and reformulate raw data into PRES datasets and tasks. However, we note that many benchmark papers, e.g. OGB (Hu et al., 2020), TGB (Huang et al., 2023), and TGB 2.0 (Gastinger et al., 2024), also focus on curation rather than creating new data. We believe this does not lessen the value of our contribution to the dataset and benchmark track.

Official Comment by Reviewer wgDA

Official Comment by Reviewer wgDA 02 Aug 2025, 02:41 (modified: 02 Aug 2025, 02:42)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=lnjd1X1Vg8)

Comment:

Thank you for your detailed reply. However, some concerns still remain:

1. For dynamic graphs, interaction information can be stored in edge features, while for event sequences, relational information can also be described through event features.
2. I was hoping to see more dynamic graph baselines. Nevertheless, the supplementary experiments partially address this concern and make a point.
3. The data curation process somewhat diminishes the overall contribution of the work.
4. While the rebuttal provides some clarification regarding the data processing pipeline, the revised version should include more detailed descriptions.

Based on my overall assessment of the paper and the rebuttal, I have decided to raise my score to 3.

Mandatory Acknowledgement by Reviewer wgDA

Mandatory Acknowledgement by Reviewer wgDA 02 Aug 2025, 02:42

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Mandatory Acknowledgement: I have read the author rebuttal and considered all raised points., I have engaged in discussions and responded to authors., I have filled in the "Final Justification" text box and updated "Rating" accordingly (before Aug 13) that will become visible to authors once decisions are released., I understand that Area Chairs will be able to flag up Insufficient Reviews during the Reviewer-AC Discussions and shortly after to catch any irresponsible, insufficient or problematic behavior. Area Chairs will be also able to flag up during Metareview grossly irresponsible reviewers (including but not limited to possibly LLM-generated reviews)., I understand my Review and my conduct are subject to Responsible Reviewing initiative, including the desk rejection of my co-authored papers for grossly irresponsible behaviors. https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/ (https://blog.neurips.cc/2025/05/02/responsible-reviewing-initiative-for-neurips-2025/)

Replying to Official Comment by Reviewer wgDA

Official Comment by Authors

Official Comment by Authors 03 Aug 2025, 17:52

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We thank the reviewer for the follow-up discussion and for taking the time to engage with our work. We appreciate the constructive feedback and the decision to raise the overall score. We would like to address the concern as follows:

Problem formalisation

We agree with the reviewer's assessment that one way to model a dataset containing both personal and relational events is to abstract out one type of event and convert the events into features. For example, the sequence of a user's personal events can be converted into aggregated node or edge features, which can then be incorporated into a graph model. Similarly, the neighbor information encoded in a user's relational events (graph) can be converted into neighbor-aggregated event features. In the context of our example of user activities on an e-commerce website (Figure 1), this strategy may work as follows:

1. (Scenario 1) To abstract out the personal events of User A, we may do the following: In the relational event of "Sending gift money to User B" (can be easily encoded in a dynamic graph formulation), we construct features that aggregate personal events (cannot be easily encoded in a graph), such as:

- How many login events occurred before this gift-send event.
- How many product purchase events occurred.
- etc

2. (Scenario 2) To abstract out the relational events of User A and feed them into a sequence model, we may create an event representing the relational action of sending money, but encode the information that User A is sending money **to User B** by combining features from both User A and User B.

However, this modeling strategy has several downsides and limitations.

1. The approach requires a manual process of feature engineering to decide which features are relevant when abstracting out or collapsing one type of event (either personal or relational) into a set of features.
2. When we abstract out one type of event into features, there is a potential loss of information during the process. For example, abstracting out personal events in Scenario 1 above results in the loss of detailed sequence information; whereas abstracting relational information in Scenario 2 leads to the loss of higher-level relations (such as two-hop connections).
3. This strategy also prevents us from taking advantage of recent advances in machine learning that focus on learning directly from raw data or events, rather than relying on manual feature engineering.

Our vision for this work.

In this paper, we aim to demonstrate the need to incorporate both sequential and relational information when modeling user events, an aspect that is often overlooked by the community. We do this by curating raw user event datasets, processing them, and constructing prediction tasks. We then show that even relatively simple models that utilize both types of information can outperform more complex models that rely on only one.

However, this is not our ultimate goal. If simple models can perform well by leveraging complete information on user events, this opens the door for exploring more complex approaches, such as end-to-end models, that utilize both raw sequential and relational information. We hope that our work can **serve as a catalyst for developing end-to-end models that integrate sequential and relational information**.

Related baselines

We thank the reviewer for clarifying that the additional requested baselines were dynamic graph models rather than base static graph models. We included TNCN as an additional baseline in our experiments, as it is one of the leading models on the TGB leaderboard. We are happy to include additional dynamic graph models in the revised version of the paper.

Data curation process

We appreciate the reviewer's opinion on this. However, we believe that our data curation contribution remains significant to the community. The area of simultaneously combining sequential and relational information in user event modeling is still underexplored. Previous benchmark datasets on user events tend to focus on either a sequential-only or relational-only formulation. We hope that our work will bring more attention to this integrated modeling perspective and support the development of a new family of advanced models that jointly capture the sequential and relational aspects of user event data.

Data processing information

We thank the reviewer for raising concerns about the description of the data processing stage in our responses. We have provided some details about our data processing, including an example using the Brightkite dataset. We also included a link to the actual script used in the data processing, which we will release to public. However, we realize that our description may still not be sufficiently detailed. We would appreciate any additional feedback from the reviewer on specific aspects that should be further clarified. We would be happy to include these additional details in the data processing section of our revised paper.

Replying to Official Comment by Authors

Official Comment by Reviewer wgDA

Official Comment by Reviewer wgDA 05 Aug 2025, 04:06

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for your further clarification. I will decide my final evaluation during AC-Reviewer discussion phrase.

Replying to Official Comment by Reviewer wgDA

Official Comment by Authors

Official Comment by Authors 05 Aug 2025, 22:19

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

We sincerely appreciate the reviewer's time and effort in reviewing our paper, engaging in active discussion, and offering valuable suggestions. If there are any other concerns or questions, we would be happy to provide any additional information. Thank you.