Journal of Visualization and Interaction
JoVI review snapshot created at Jan 27, 2026 08:55.

Issue #18 (open): [DECISION] GLaDOS: Graph Layout Algorithm Benchmark Datasets for Open Science â€“ GD Benchmark Sets

@codementum on Jun 27, 2025 20:00: [opened]

@codementum on Jun 27, 2025 20:00:

Conflicts of interest

- ☒ I declare that I have no known conflicts of interest with the authors.

Reviewed version

f41e831

Reviews summarized

- #11
- #12
- #17
- #15
- #14

Meta-Review

Reviewers generally agree that this paper represents an important contribution for researchers working with graph datasets. Reviewers particularly liked the systematic approach, the interactive features, and the potential for longer-term impact. At the same time, reviewers raised concerns and made suggestions about these and other issues.

Overall, the paper seems well-positioned for minor revisions, though there are several points which may require additional review (e.g. clarifying issues about the taxonomy).

The authors should review individual reviews #11, #12, #17 for specific items to address. These can even be referenced using GitHub features as changes are made. Here also is a summary of concerns that came up across several reviews, with some additional thoughts and suggestions:

- Justifying or re-framing the taxonomy (relevant reviews: #17, #11). In particular, #17 raises several critical points on the framing and definitions surrounding the taxonomy. One option may be to reframe-- i.e. taxonomy has a specific meaning in certain research communities-- is this paper trying to propose one? Might a different term be used? Or is the goal to develop a useful set of terminologies and categorizations to support the dataset project at this early stage? Another option might be to lean into the taxonomy term and more rigorously develop and define the proposed taxonomy.
- Long term maintenance of GLaDOS (relevant reviews: #11, #12, #17). All reviews raise this issue in some form. There appears to be tangible value and potential in the proposed GLaDOS approach. For example, OSF was seen as a positive for its commitment to longevity. However, reviews surfaced multiple specific concerns, such as how to update, how to broaden (e.g. other communities). The authors might consider updating with some additional discussion on possible maintenance models. (One potential suggestion here: there may be a useful community-driven model to be developed, i.e.

using the issues and pull-requests features of GitHub and similar platforms for the community GLaDOS supports. The IEEE VIS website model comes to mind.)
- Features of the GLaDOS interface (relevant reviews: #11, #12, #17). Reviews raised specific requests for additional filtering criteria or features for the website interface. I would suggest the authors either: 1) consider these and implement them where possible or 2) or document them and provide explanation for why they may be technically out of scope or otherwise not in line with how GLaDOS proposes to support navigation of datasets. (Again, Issues for features might be a nice way to open up dialogue about these. Certainly not required, but potentially useful for more focused discussion if helpful.)

Authors and reviewers are invited to discuss changes and seek clarifications as supported by the platform and reviewing process.

Decision

Minor revisions: this paper requires some smaller changes, after which I am confident I would be able to endorse it.

@picorana on Jul 01, 2025 11:12:

Thank you for the review! I will be working on the changes soon.

@picorana on Aug 04, 2025 18:33:

Thank you, reviewers, for the very detailed reviews! I did my best to address all of the comments. I hope I managed to satisfy the requests.

Accessibility: I removed the citation tooltips as requested in Issue #15. I tested the website with macOS voiceover, as it has been done in Issue #15, and the navigation does not get stuck in the bars of the barcharts for me (using control+option and directional keys to navigate the page). Perhaps I am missing how to replicate the problem?

Oh, we also started using uv for simplifying the setup as mentioned in #13.

Changes to the website: The website is now a little faster when loading - before, users needed to wait some time for the contents to be loaded. I achieved this by removing a dependency, which wasnâ€™t really used much and was somehow slowing down everything. I also added an additional page - accessible from the top of the main index page, called about.html, which contains additional explanations, information on how to contribute, citation info...

Search function: Issue #11 mentions, as a feature to be added, a search function. I added a search bar both in the website and on the paper, where datasets can be searched by title and feature.

â€œTaxonomyâ€ wording:

> (Issue #17) "This work proposes an overarching taxonomy of datasets", but I feel it is a stretch to call it this. There are a number of inconsistencies and questions about the categorisation, as described below. The classification should be made more thorough and precise, or this claim reduced.

I agree: There really was no reason to call this a taxonomy, and it is present in only one sentence in the whole paper. I changed the wording to â€œworking classificationâ€.

Maintenance (Issue #12, Issue #17): I created an issue template on github to submit requests. For maintenance, we use Notion to keep track of the datasets and to collaboratively edit the CSV files. While Notion is a proprietary platform, all the data remains in plain CSV format, which means we can easily migrate to other tools if needed. The use of Notion is just a convenience, nothing is locked into it.

I expanded a section in the body of the paper to include this: â€œMaintenance plans and contribution to the repository: Contributions to the dataset collection (corrections, integrations, replacement) are most welcome â€” and strongly encouraged. However, to ensure data quality and avoid accidental overwriting or inconsistency, we donâ€™t allow direct edits to the files by everyone. Instead, there are two main ways to contribute: Pull requests: If you prefer to fill in all the information yourself, you can submit a pull request directly to the repository. The data for both the papers and the datasets is stored in CSV format, available here. GitHub issues: If youâ€™d rather just point us to a new dataset or share some

additional info (e.g., missing metadata, clarifications, or links), weâ€™ve created an issue template to make that easier. Weâ€™ll then take care of adding the dataset and filling in whatever information we can find or infer. Even without external contributions, we actively monitor the space and try to keep the repository up to date as new datasets emerge. And if none of the options above work for you, feel free to just reach out to us â€” weâ€™re happy to handle things more informally as well. That said, itâ€™s worth noting a clear limitation: this collection does not aim to be exhaustive. The starting point for the dataset list was a literature review covering a few hundred papers, which means itâ€™s entirely possible that some benchmarks were missed â€” especially if they werenâ€™t cited often or were introduced in more obscure venues. For this reason, contributions from the community are especially valuable to help fill in the gaps and keep the resource as useful and complete as possible.â€

Some related comments in the reviews are addressed here:

> (Issue #17) The future maintainability of the work is unclear. That is, there is no commitment or plan in the article of the future maintainability of the archive as new graph data sets are published. It would be good to see a plan for this, whether it is the authors or as an open project. If the later there might be a need to document the processes for maintenance to allow others to update the GLaDOS website.

and

> (Issue #17) The article makes it clear that all datasets are sourced from the review of literatures, but the GLaDOS website doesn't give any details of where they were sourced, other than the "Benchmark datasets" category saying "These are collections of graphs that have been frequently used in graph drawing papers" which is then not listed under the other categories. There should be a sentence or two on the website that explains all datasets came from graph drawing literature.

I added a section in the â€œabout.htmlâ€ part of the website, and expanded the invitation to contributing in the paper, as well as pointing to the right resources containing the data and creating a github template for contributing.

Discrepancies between the website and the paper:

> (Issue #17) The descriptions on the website are clearly hand-authored and are inconsistent in the presented information.

The hand-authored descriptions in the paper were intentionally written in a more colloquial and narrative tone, to make the content feel like a cohesive text rather than a series of catalog entries. This choice was made to support readability and discussion, especially for a publication format where we wanted to highlight patterns, nuances, and commentary rather than simply list dataset fields.

That said, all the factual information present in these descriptions â€” such as number of graphs, structural properties, and other metadata â€” is systematically reflected in the underlying data and is accessible through the structured CSV files and website. We understand the importance of consistency, and while the prose may vary to suit narrative flow, the core dataset information remains complete and aligned across both formats.

Clarification on the classification of the datasets:

> (Issue #17) Some of the Datasets are classes of graph rather than a single data set. For example "Social Networks" is from 4 data sources. Does this mean all the papers that use this use all 4 or just 1. If the later, this makes the presentation of the most-used data sets questionable with Social Networks being number 4.

Most papers only use one of them, not all four. We grouped them to reflect how often â€œsocial networksâ€ as a class of graphs are used in layout evaluations, but we see how this may inflate the appearance of that group in the ranking.

To address this, weâ€™ve now added a note in the ranking section to explain how grouped categories were handled. We also revised the language to make it clear that the ranking reflects grouped usage, not necessarily a single, unified dataset.

> (Issue #17) The difference between "Uniform Benchmark" and "Established Network Repository" is not clear, especially as it related to subset collections.

and

> (Issue #17) The article says "Uniform Benchmark datasets" "aims to provide a general overview of the performance of a graph layout algorithm by testing on a large amount of graphs varying in size and density, rather than focusing on a specific type of graph". It is not clear what "specific type of graph" means. For instance, it could be argued that all graphs in the Storylines dataset are a specific type of graph.

By **Uniform Benchmark*** ***we mean datasets that were explicitly curated and released together for the purpose of evaluating layout algorithms (e.g., with common format, goals, or metrics in mind). In contrast, an*** ***Established Network Repository** refers to larger, more general-purpose archives like SNAP or KONECT, which collect a wide variety of networks, often across domains, and not specifically for layout evaluation. Subset collections taken from these repositories fall into the second category unless they were later re-released in a more curated form for layout evaluation â€” in which case we list both (and cross-reference them in the dataset notes). Weâ€™ve clarified this in the text.

> (Issue #17) Why categorise the North DAGs and AT&T graphs separately, and then in the North NAGs description, just have See AT&T". Given the focus on reproducibility, It would make more sense to have this as an alias for the AT&T dataset and combine them in the listing.

I think it would be relevant to keep this information explicit (which paper calls a certain dataset with which name) - also, if any author is looking specifically for â€œNorth DAGsâ€ (because it was perhaps mentioned on a paper that called it with this name), it is a good way to redirect them to the other version of the dataset.

> (Issue #17) For the "Established Network Repositories" the article says "we do not include here any storage of the data (which would be redundant) or report statistics on them" but provided information (size information, min/max nodes, node distribution and summary charts) would be useful since that allows people to select the appropriate data set. If the issue with generating stats is that they could becoe outdated if the collection is added to, you could say the information was based on a date and then potentially update these as part of the maintenance of GLaDOS.

So this is a choice we had to make. Arguably, these other network repositories are doing a job that is similar to GLaDOS, only not very focused on benchmarking layout algorithms. It is not really my intention to re-do their work, and neither it is to re-upload their data, as clearly there are active groups of maintainers working on them. They do report their own statistics, they did their own work, and itâ€™s not my intention to step on their toes by replicating their work. We only cared about storing datasets that are clearly not currently cared for.

> (Issue #17) The KnownCR set lists the "Known Crossing Number" tag, but is this value included in the data (graph downloads) for each graph in the dataset?

No, for the same reason as above.

> (Issue #17) The article says "One more of such collections is Konect. At the time of writing, though, the website for Konect has been down for a while. Both the data and the website are still accessible through web archiveâ€”thus we do not consider this a lost collection." Why is the Konect dataset not included in the collection then?

Luckily we found that Konect is back online! I added it back to the list of datasets. I did not include in the original list because we have no papers actually referencing Konect as a source for their dataset. It is a bit of a choice I had to make: do I include a dataset even if it is not used in any of the papers I include in the ones I collected? I am aware of it, but it did not come up using the same methods we used for the rest of the collections. Anyways, itâ€™s really valuable as a collection, so I added it back. However, I do not have any examples of how it was used in papers.

> (Issue #17) For a dataset like "Car Features" where the features are not known, the dataset is unavailable and the author won't disclose the origin of the data, is there value including it at all, except maybe in a list of papers that fit this category (I imagine there might be many).

We still think itâ€™s important to document the existence of this, not as usable benchmark, but as part of the landscape of datasets that have been referenced in layout evaluation work, especially when they appear in published papers.

Clarification on â€œcustom-madeâ€ datasets, reconstructed datasets (Issue #11):

It is important to make it clear that we did not attempt reconstructing any dataset - we only collected them as-is, when we could. I clarified this in the paper.

> (Issue #11) Discuss Potential Bias in Dataset Selection â€“ Since some datasets were reconstructed, it would help to include a short note on how the selection process might introduce bias and what was done to minimize this. We did not reconstruct any datasets â€” all of them were taken as-is from existing sources. We only worked towards finding datasets, never reconstructing them.

The one thing that might introduce some bias is how the datasets were collected in the first place. We mostly relied on literature search and snowballing, so of course we canâ€™t claim to have found everything that exists. Some datasets may have been missed simply because they werenâ€™t cited often or were harder to find. We now mention this in the paper to make it clearer, and we added a note that the repository is open to community contributions to help fill in any gaps.

> (Issue #11) Clarify Data Consistency Checks â€“ There is no clear mention of how the authors ensured data accuracy and consistency across different sources. A brief explanation would strengthen transparency.

We donâ€™t do aggregation between different sources â€” we mostly report the data as it appears in the original sources, all of what we could find. Thatâ€™s also why we include descriptions from the literature directly: if two papers describe the same dataset differently, we donâ€™t try to resolve the contradiction, we just show both. This way, we avoid introducing our own interpretation and let readers see how the dataset has been presented in different contexts. Weâ€™ve now clarified this point in the paper as well.

> (Issue #17) It also talks about sub-categories of Replicable vs. Reproducible vs non-replicable. Where does this information appear in the GLaDOS website? This is not clear.

The GLaDOS website only hosts datasets that we found in their entirety. We never attempted reproducing a dataset â€” especially not the ones that didnâ€™t give enough details to be perfectly replicable, because this could introduce biases, issues, imprecisions. Because of this, we do not deem really relevant to have this information on the website, as there is no mention of datasets that are not perfectly replicable. The GLaDOS website does not host anything that we classified as â€œcustom-madeâ€.

Highlighting the limitations of the collection (Issue #12)

We do recognize that some papers might have been overlooked. However, as per every survey, we need limits. Anyways, we streamlined a bit the way in which people can contribute, by having a github issue template and better explaining how maintenance will be handled (see previous section, â€œMaintenanceâ€).

Clarifying differences with W. Hu et al. (Issue #17)

> (Issue #17) When mentioning the work of W. Hu et al. having a different focus, the article should explain why it is not relevant.

Expanded explanation: â€œThe Open Graph Benchmark collection from W. Hu et al. (2020) is also worth mentioning. It provides an important infrastructure for evaluating machine learning methods on graph-structured data, including datasets, tasks, and evaluation metrics, but it is not focused on layout quality or human-perceived readability of graph visualizations. Our work complements such efforts by specifically targeting datasets designed to evaluate the perceptual and aesthetic dimensions of layout algorithms, which are not addressed in W. Hu et al. (2020).â€

Clarifying how we found some unavailable datasets

> (Issue #17) The article says they looked "into internal storages of research groups". Please clarify what this means. Does it mean the authors contacted someone else from their previous group, if the author was no longer there and didn't have the data themselves, and they looked in the internal storage?

Yes, thatâ€™s more or less what we meant. Weâ€™ve now clarified the sentence in the paper. In a few cases, when datasets were not publicly available, we reached out to the authors directly, and if we had

access (for example, by being part of the same institution or having contacts in the group), we checked whether the data was still stored somewhere internally â€” like in shared drives, old project folders, or archived backups. This sometimes helped retrieve datasets that were not accessible online anymore. Weâ€™ve reworded the sentence in the paper to make this clearer. â€œWe also reached out to colleagues in other universities who we knew had worked with certain datasets in the past, and asked if they could check their internal storage â€” for example, shared drives or old project folders â€” to see if the data was still available.â€

File formats:

> (Issue #17) The section on file formats (3.1) is unclear. It says "We chose to convert and store several of the datasets in a uniform JSON representation". Why only "some"? It says "we have also converted and made available all graphs in three additional commonly-used formats: GraphML, GEXF, and GML". Why have "multiple accessible formats"? Do these all have exactly the same information, or do some lack the extra information ("timestamps, labels, or belonging to a clusters, and edges having weights"). The article and website should make clear if there is a master format for everything and then convenience formats and what information they each contain. Otherwise people could end up comparing graphs that are actually different (for info like clusters) when thinking they are the same.

It was specified on the paper that we chose JSON as the â€œmainâ€ format. All of the other formats are provided as a courtesy and contain the same information (metadata, additional attributes). I updated the description to be explicit about the fact that all of them are equivalent.

Other minor comments:

> (Issue #17) "the graph structure is very difficult to piece together" This could be explained. Is it because it is in an uncommon format, or that the source data is something like Twitter posts.

This was about uncommon formats. I specified it in the paper.

> (Issue #17) The figure (data collection process) at the beginning of Section 3 should be labelled and have a caption.

Done. It actually had a caption, but was missing the â€œFigureâ€ label.

> (Issue #17) "the original data we downloaded when the files were small enough to be uploaded to GitHub". What does this mean? Why does it matter?

It did indeed not matter. I removed it.

> (Issue #17) In Section 4.2 "Random Generation"The article says "The list of features to take into account to claim that a synthetic graph is comparable to another one would be long, and perhaps out of the scope of this publication. These are just a few examples of what could be relevant:" What follows looks like a list produced by GenAI with no explanation of where this came from (at least I got similar result crafting a quick prompt from the text above). I question what this list contributes to the paper. If it is to be included, it should be shortened and properly justified.

So sorry, that list was actually the result of a brainstorming session and not made with AI. As you can see from the amount of typos you spotted, AI was used very little for this paper (: I have replaced the list with this statement:

â€œComparing two synthetic graph collections meaningfully is not straightforward. Without shared generation procedures or metadata, itâ€™s difficult to claim that they represent comparable input conditions for layout evaluation. A number of structural and contextual features may affect comparability, including, for example, graph size (nodes and edges), density, and distribution of components or motifs. Even small changes in these properties can have a significant impact on layout results. While a full discussion of all relevant factors is beyond the scope of this paper, we highlight this issue to caution against assuming that two synthetic datasets are interchangeable just because they were randomly generated. A more careful analysis of their structural characteristics is often necessary.â€

> (Issue #17) References should be given for "ErdÅ‘s-RÃ©nyi", "BarabÃ¡si-Albert" and also the sentence beginning "Conversely, the BA model produces scale-free networks with... "

References added!

> (Issue #17) In Section 5, the difference between gray and blue dots in Figure 8 is almost impossible to distinguish. Please use more visually distinct colours or a different mark.

Color was changed to red.

In addition to these items, I fixed all the typos mentioned in the reviews. Thank you for finding them!

@domoritz on Sep 01, 2025 22:07:

I re-checked the voice over navigation in https://www.journalovi.org/2024-dibartolomeo-benchmark/. The annotations for the charts are pretty bad as the individual bars are just images and the labels of the axes get read out. Ideally the bars should say the label and value. I can skip over the group by using control+opti0on+shift and then left and right. Those skip over the groups. But the content of the groups is still mostly noise and should ideally be hidden from the screen reader. Please try to impove the screen reader experience but I think this may also be because of the library you use. Maybe file an issue with them.

Thanks for making the other fixes!

Either way, because the groups can be skipped, I consider this not blocking for acceptance. So I approve the paper.

@floe on Dec 17, 2025 08:14:

Hi everyone, since all reviewers now endorse the paper, I'm preparing the actual publication in JoVI. I noticed a few minor things to address (/cc @picorana):

- ☒ some references have the author name listed as "---"
- ☒ the "under review" box on top of the article can be removed
- ☒ JoVI provides a backup PDF for live articles, can you generate one from Quarto directly?
- ☒ pick a cover picture for the publication website

Thanks!

@floe on Jan 09, 2026 09:53:

Hi everyone, happy new year! Just a reminder that there are still some minor copyedits outstanding (see previous post) before we can officially publish this on the JoVI repository... (/cc @picorana).

@picorana on Jan 09, 2026 10:50:

Thank you for the reminder! I will close down all the minor issues in the next couple of weeks, thank you!

@picorana on Jan 22, 2026 23:07:

Hello!

My last commit addresses the changes that were requested.

The only thing I had issues with was generating a PDF: quarto will not render the result of ojs blocks in a PDF. I found this discussion about it in quarto issues: https://github.com/orgs/quarto-dev/discussions/1909 As a result, the PDF looks comically ugly and pretty useless, without the charts.

The only alternative I could find was to print the HTML-rendered page to PDF, which I did here: PDF version

> pick a cover picture for the publication website

Is this one okay? picture

Will the publication be indexed in Google Scholar? Or should I maybe consider uploading to arXiv?

Thank you so much!

@floe on Jan 25, 2026 11:07:

Thank you @picorana, looking good! Yes, JoVI publications are indexed by Crossref and consequently also show up in Google Scholar. I'll push the remaining changes to our publication system at AAU early next week.