
Issue #15 (open): [REVIEW] Gatherplot Review 2

@facet-fan on Sep 20, 2023 21:59: [opened]

@facet-fan on Sep 20, 2023 21:59:

Conflicts of interest

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

Reviewed version

0ee8b81

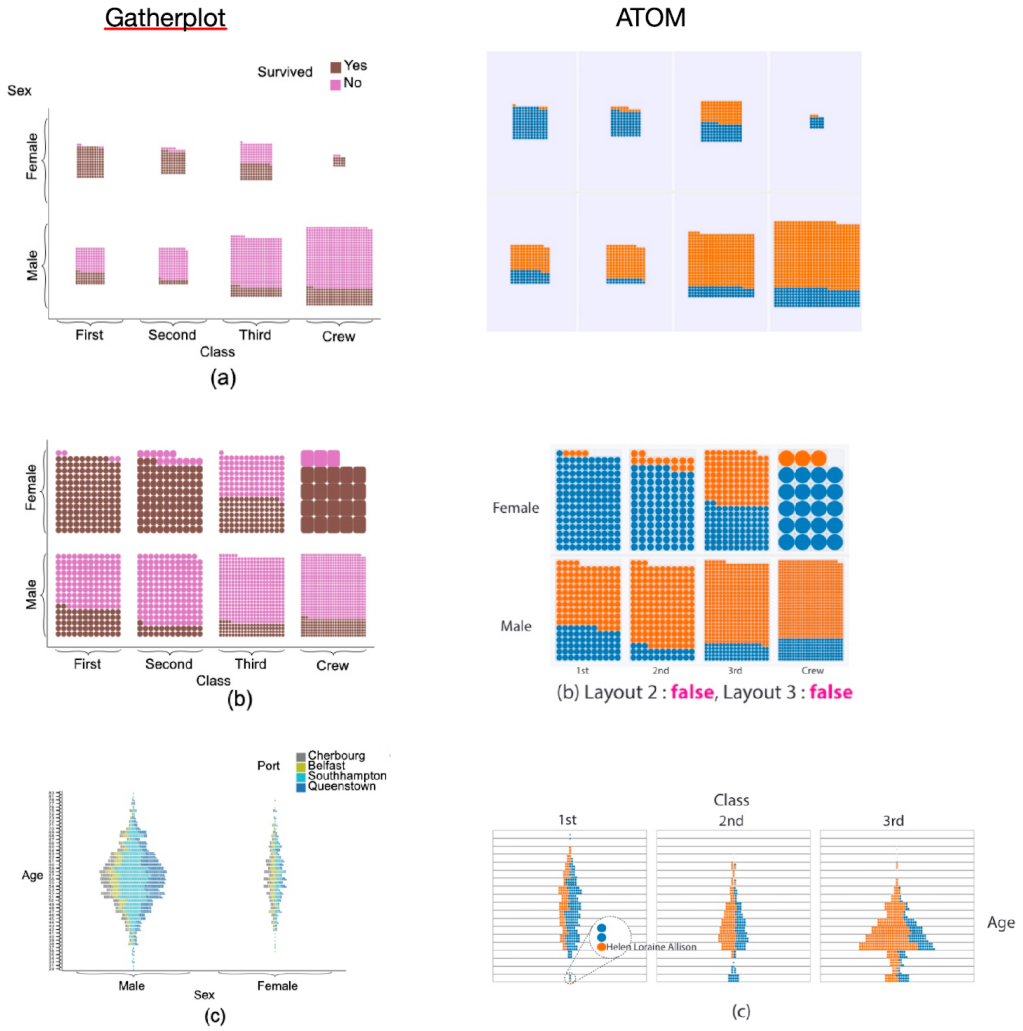
Review

Summary

The authors present Gatherplot, a unit visualization technique that packs marks within p
overplotting. I think Gatherplot is a useful technique for solving a common problem, but
more work in refining the contribution and improving the validity of the design and the s

Knowledge

I hope that the authors make their knowledge contribution clearer to make this paper a s
submission.



The expressiveness of Gatherplot appears to be a subset of the ATOM grammar [Park et a knowledge added is not immediately clear. The two papers even share similar figures (se What the authors can do is to more explicitly compare Gatherplot and ATOM and outline added (like more details on algorithms, design decisions, and iterations) or how they are ATOM is only cited in passing and the comparison is hard. Alternatively, the authors can Gatherplot evolved into ATOM if that is the case. Then the knowledge contribution might designing a visualization grammar".

In addition, this paper is on the older side. A version of it is a technical report from 2016 the most recent references listed were from 2018. What new and relevant knowledge hav community) gained about unit visualizations and remedying over plotting for the last 5 y

Given how similar Gatherplot and ATOM are and how old this manuscript is, the authors more context on Gatherplot to make the paper useful for readers today.

This paper also has a user study for evaluation. I have outlined some issues and ambiguit addressable with open materials, see Validity section below. Without revisions, the user s towards sound knowledge yet.

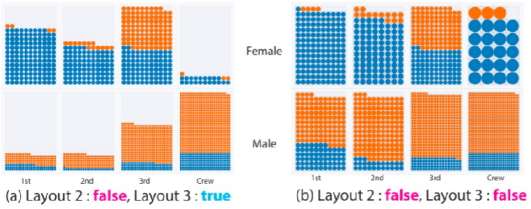
Validity

Design

The sections discussing the design of Gatherplots could be improved with more details ar

The optimal bin size, an important feature of Gatherplot: - The authors list "a heuristic to optimal space usage" as a contribution, based on "spatial accuracy and legibility". I can u

legibility---the individual dots/marks should look distinct from each other. But the author explain what spatial accuracy is and how it is applicable in this context. In Figure 6, the a has better spatial accuracy. I assume that this accuracy means the deviation from the x-a deviation should fall within the appropriate x-axis segment from the gather transform, s major issue to optimize for? My concern is that there looks to be aliasing in Figure 6 desp "accurate"--- some dots appear to have been sorted into small bins/rows where the data i might want to consider how their bin size choice affects the perception of the data distri the validity of the bin size algorithm, I hope that the authors can summarize their appro function called `getOptimalBinSize()` from https://github.com/intuinno/gatherplot/ blob/5d4e902d262986219e50b9624694789f0bb5f281/app/scripts/directives/gatherplot.js#I can turn it into a few sentences. - A minor comment: Wilkinson proposed a single bin siz "a well-written dot plot program should automatically down-size dots when extreme ove (Wilkinson, 1999), which is what the Gatherplot algorithm does from my rough reading o authors should address the fact that this variant of Gatherplot (categorical + continuous v of dotplot (also called beeswarm). If applicable, the authors should consider citing more dotplot implementations and comparing them to how Gatherplot works. - Vega-Lite, Obse `smooth` option that mitigates aliasing https://observablehq.com/@uwdata/dot-plots - ggpl that automatically picks the bin size depending on the aspect ratio https://osf.io/2gsz6



(a) Layout 2 : **false**, Layout 3 : **true** (b) Layout 2 : **false**, Layout 3 : **false**

Fig. 8. Sharing can be applied hierarchically. Here the *Titanic* dataset has been faceted by gender and passenger class. In Figure 8, every facet shares the size by setting the size sharing property of "layout2" and "layout3" as true. This yielded a unit bar chart where every dot size is same and the size is adjusted such that the most crowded facet can fill the assigned space. However, (a) shares size only in layout2 that the unit will be the same size among the class but not across genders. This is in contrast to (b), where sizes are independent of gender and class, meaning that every unit will be scaled up to fill their subcontainer.

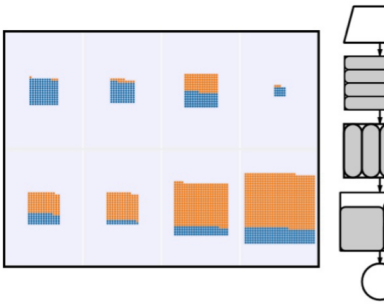


Fig. 10. Unique visualization generated using A of the *Titanic* were faceted according to their class (horizontal), and then each unit was colo passengers survived).

The choice to use `Pack; Size : Count, Shared` instead of a bar chart layout - Figure 1 paper is a Gatherplot, with the ATOM spec `Pack; Size : Count, Shared`. With the sar is a version without the `Pack` and resembles a faceted stacked bar chart. I hope that the a they did not make the bar chart layout the default in Gatherplot: bar charts (even when f encode values by position on a common scale, which *should be* more perceptually accura with area (what Gatherplot does). Also, the subgroups of dots will share the same width/h dimension, making sub-group comparisons easier.

Misc comments

- In Figure 5, if I select `x-axis = Name`, the x-axis only has one segment "VW rabbit c are many more car names in this dataset, visible when selecting `color = Name`. Is t choice?
- The "Using the Gather Transformation" section talks about parallel coordinates (wit instead of scatterplots (2D). I find it difficult to follow without visuals or context, and add a figure or remove the section if it is not closely relevant to the rest of the paper.

Validity of the evaluation

The authors evaluate the "effectiveness of Gatherplots" "with categorical vs. categorical v crowd-sourced experiment has a simple design (a positive) and the authors applied appr tests. However, there are inconsistencies in the analysis and a lack of details, part of whic addressed by more open materials on the experimental stimuli and analysis code.

Study design - The authors should provide more details on the exact wording of the task

the question (number input or multiple choice etc). Were participants expected to count (many) to answer the questions? Does incorrect mean that e.g. the participants input 10 a answer is 9? Perceptual studies often characterize accuracy through continuous variable measure in Cleveland & McGill, 1984), so do the authors have a particular reason for enc correct/incorrect instead of a continuous error measure? I think information on how the designed can help readers assess the validity of this study.

The "both" condition - The authors test a "both" condition where the participants can to absolute and normalized mode. However, this condition is not included in the hypotheses discussed in the accuracy analysis. The authors only report that the `Absolute * Relati` significant without follow-up analyses to interpret the interaction. Also, for the completic people spend the most time in the "both" condition because they toggled between the mo Casual users don't often take advantage of interactivity in visualizations. The authors mig whether the toggle is worth it given the completion time and accuracy performance.

Analysis - My main issue is that I am unable to reproduce the reported statistics ($\chi^2 averages while the analysis should be pretty straightforward. Again, the authors should k about how they performed the analyses and provide analysis code given that Jovi advoca practices. - For example, for Task 1, jitter condition, the authors report the average compl 44.26s, but my analysis returns ~ 36.5s. Figure 12 also shows a mean < 40. The authors sh discrepancy. - The authors should describe how repeated measures are handled in their l - The authors state that "gatherplots enable people to assess data distribution more quick but I could not find support for this in the analysis. The authors should double-check and descriptions.

Interpretation of the results - The authors should consider offering possible explanatio and design recommendations. Could gatherplot be better than jittering because the dots a layout and therefore easier to glance at or count? How should we use gatherplots given t results? Having interpretations like these on top of a list of statistics can make this paper informative for future research.

Misc writing

- The length of this paper is appropriate for its content, though I did suggest a few pla authors could elaborate more.
- Small things
  - Section 1: "...overlap is known as (or) in visualization"
  - Two paragraphs start with "Finally" in the "Data-aware Methods" section.

Openness/Transparency

As a condition for acceptance, I hope the authors can provide the stimuli, data analysis s applicable materials per the Jovi requirements https://www.journalovi.org/author-guide.l requirements.

The CSV file for the experimental data matches the number of participants, conditions, a described in the paper. The README describes the data columns well. I did not find the a OSF link.

Submission categories

- ☐ Registered Report
- ☐ Replication Study
- ☑ Empirical Research - Quantitative
- ☐ Empirical Research - Qualitative
- ☑ Systems or design research
- ☐ Commentary
- ☐ Systematic Literature Review

Suggested outcome

Major revisions: this paper requires substantial improvements that I will need to re-revie
whether or not to endorse it.

Requested changes

- Refine the knowledge contribution in the context of ATOM and potentially newer lite
- Clarify the bin size algorithm and justify the packing layout decision
- Provide experimental stimuli and data analysis scripts
- Resolve analysis issues and elaborate on how analysis results translate to the practic
  gatherplots.

ORCID

No response

@codementum on Dec 20, 2023 20:33: [referenced from #[DECISION] Gatherplots 20-Nov-2023]

@nickelm on Mar 24, 2024 20:05:

> The authors present Gatherplot, a unit visualization technique that packs marks wit
> and avoids overplotting. I think Gatherplot is a useful technique for solving a comm
> this paper requires more work in refining the contribution and improving the validi
> and the study.

Thanks for the feedback. I have tried to address it exhaustively.

> I hope that the authors make their knowledge contribution clearer to make this pap
> submission.
>
> The expressiveness of Gatherplot appears to be a subset of the ATOM grammar [Par
> so the knowledge added is not immediately clear. The two papers even share similar
> figure above). What the authors can do is to more explicitly compare Gatherplot and
> outline new knowledge added (like more details on algorithms, design decisions, an
> how they are different. Currently, ATOM is only cited in passing and the comparison
> Alternatively, the authors can reflect on how Gatherplot evolved into ATOM if that is
> the knowledge contribution might turn into "lessons in designing a visualization gra

Revision: We have clarified the distinction in a direct comparison at the end of the subse
aware Methods".

> In addition, this paper is on the older side. A version of it is a technical report from 2
> ATOM), and the most recent references listed were from 2018. What new and releva
> have we (vis community) gained about unit visualizations and remedying over plott
> 5 years?

Revision: We have improved added missing references, and will add further ones for the

> Given how similar Gatherplot and ATOM are and how old this manuscript is, the aut
> provide more context on Gatherplot to make the paper useful for readers today.

This is a fair point.

> This paper also has a user study for evaluation. I have outlined some issues and amb
> addressable with open materials, see Validity section below. Without revisions, the u
> count towards sound knowledge yet. The sections discussing the design of Gatherplo

improved with more details and justification:

> The optimal bin size, an important feature of Gatherplot: The authors list "a heuristi
> sizes for optimal space usage" as a contribution, based on "spatial accuracy and legil
> understand legibility---the individual dots/marks should look distinct from each othe
> authors might want to explain what spatial accuracy is and how it is applicable in th

Revision: We have added definitions to both legibility and spatial accuracy to the beginn
"Managing Continuous Variables" subsection.

> In Figure 6, the authors state that (a) has better spatial accuracy. I assume that this a
> the deviation from the x-axis tick.

No, it is the deviation from the y-axis tick, not the x-axis tick (since this is a one-dimensio
position actually does not mean anything). We hope that the new definition of spatial acc
clarify this point.

> But any deviation should fall within the appropriate x-axis segment from the gather
> deviation still a major issue to optimize for? My concern is that there looks to be alia
> despite being more "accurate"--- some dots appear to have been sorted into small bir
> the data is dense. The authors might want to consider how their bin size choice affec
> perception of the data distribution.

Yes, this is exactly our point: Figure 6b does indeed yield aliasing, i.e. lower spatial accura
smaller chart size precludes the dots from being made smaller (because that would viola

Revision: We have added a sentence noting this as an inherent weakness of unit visualiz

> To add to the validity of the bin size algorithm, I hope that the authors can summari
> approach. I found a function called getOptimalBinSize() from https://github.com/intu
> gatherplot/blob/5d4e902d262986219e50b9624694789f0bb5f281/app/scripts/directives
> gatherplot.js#L2253. The authors can turn it into a few sentences.

Revision: Thanks for pointing this out. We have added this iterative search to the descrip

> A minor comment: Wilkinson proposed a single bin size but also said that "a well-wr
> program should automatically down-size dots when extreme overflow occurs" (Wilk
> which is what the Gatherplot algorithm does from my rough reading of the code.

Revision: Thanks again---we have added this to the paper.

> The authors should address the fact that this variant of Gatherplot (categorical + con
> variables) is a type of dotplot (also called beeswarm). If applicable, the authors shou
> citing more recent open-source dotplot implementations and comparing them to hov
> works.

This is a good point echoed by earlier reviews.

Revision: We have added a discussion of stripcharts, stripplots, beeswarms, and swarmp
work.

> Figure 10 of the ATOM paper is a Gatherplot, with the ATOM spec Pack; Size : Count,
> the same data, Figure 8(a) is a version without the Pack and resembles a faceted stac
> hope that the authors address why they did not make the bar chart layout the defaul
> bar charts (even when filled with dots) encode values by position on a common scale
> be more perceptually accurate than encoding with area (what Gatherplot does). Also
> of dots will share the same width/height on one dimension, making sub-group comp

Atom, as noted, is an evolution of gatherplots. We felt that combining bar charts inside a
potentially confusing.

> In Figure 5, if I select x-axis = Name, the x-axis only has one segment "VW rabbit cus
> there are many more car names in this dataset, visible when selecting color = Name.
> a design choice?

This is a bug; we're addressing it. Thanks for spotting it!

> The "Using the Gather Transformation" section talks about parallel coordinates (with coordinates) instead of scatterplots (2D). I find it difficult to follow without visuals or the authors should add a figure or remove the section if it is not closely relevant to th paper.

Revision: The parallel coordinate example was far-fetched; we have removed it.

> Validity of the evaluation The authors evaluate the "effectiveness of Gatherplots" "wi vs. categorical variables". The crowd-sourced experiment has a simple design (a posi authors applied appropriate statistical tests. However, there are inconsistencies in th a lack of details, part of which might be addressed by more open materials on the ex stimuli and analysis code.

We have tried to address this in the new revision. Our OSF repository has been updated. stated elsewhere, given the age of the study (2014?) and our student co-authors having de pastures, we have not been able to reconstruct everything.

> Study design
>
> The authors should provide more details on the exact wording of the tasks and the fo question (number input or multiple choice etc). Were participants expected to count many) to answer the questions? Does incorrect mean that e.g. the participants input correct answer is 9? Perceptual studies often characterize accuracy through continu (like the error measure in Cleveland & McGill, 1984), so do the authors have a partic encoding responses as correct/incorrect instead of a continuous error measure? I thi on how the questions are designed can help readers assess the validity of this study.

Fair point. We have found the original survey and included it. As for the individual quest used correct/incorrect and no continuous error measure. I agree that a continuous measu better.

Revision: The OSF includes a full study as a PDF. The new Figure 11 shows an example of

> The "both" condition
>
> The authors test a "both" condition where the participants can toggle between absolu normalized mode. However, this condition is not included in the hypotheses, nor is it discussed in the accuracy analysis. The authors only report that the Absolute * Relati is significant without follow-up analyses to interpret the interaction. Also, for the co measure, did people spend the most time in the "both" condition because they toggle modes (is this logged)? Casual users don't often take advantage of interactivity in vis authors might want to discuss whether the toggle is worth it given the completion ti accuracy performance.

This is a fair point. We added the condition for completeness, but did not include it in the we felt that its use was amply covered by the individual absolute and normalized modes.

Revision: We have added a paragraph to the new Discussion section on the "both" mode.

> Analysis
>
> My main issue is that I am unable to reproduce the reported statistics and averages analysis should be pretty straightforward. Again, the authors should be more explici they performed the analyses and provide analysis code given that Jovi advocates for practices.

Revision: We have added the analysis scripts to the OSF. However, as the script is interac provide a direct replication. Hopefully, the reviewer will be able to validate the correctne however.

> For example, for Task 1, jitter condition, the authors report the average completion t

> 44.26s, but my analysis returns ~ 36.5s. Figure 12 also shows a mean < 40. The autho
> resolve this discrepancy.

Well spotted, thank you. This was a mistake persisting from our very first analysis; not er
there was a discrepancy.

Revision: We have recalculated all of the averages, redone the analysis, and updated the

> The authors should describe how repeated measures are handled in their logistical r

Unfortunately, I do not have the answer to this question. Furthermore, the analysis script
available. In response, I have removed the logistic regression and ANOVA results. I have a
interpretation of the confidence intervals and effect sizes.

I have kept the Kruskal-Wallis tests for the perceived confidence metric because it is clea
that we used Bonferroni corrections to account for multiple comparisons.

Revision: We have removed the logistic regression and ANOVA results from the analysis.
added interpretation of effect sizes and confidence intervals.

> The authors state that "gatherplots enable people to assess data distribution more qu
> Abstract, but I could not find support for this in the analysis. The authors should dou
> or modify their descriptions.

Revision: Good point. We removed the "more quickly" part from the abstract.

> Interpretation of the results
>
> The authors should consider offering possible explanations of their results and desig
> recommendations. Could gatherplot be better than jittering because the dots are in a
> layout and therefore easier to glance at or count? How should we use gatherplots giv
> experimental results? Having interpretations like these on top of a list of statistics ca
> paper more useful and informative for future research.

Revisions: Thanks for this feedback. We have added a new Discussion section where we
several of these explanations and generalizations.

> Misc writing
>
> The length of this paper is appropriate for its content, though I did suggest a few pla
> authors could elaborate more.

Revisions: Agreed. We have tried to do so based on the reviewer's feedback.

> Small things
>
> Section 1: "...overlap is known as (or) in visualization"

Revisions: Thanks. This was actually due to some lingering \textit{} and \textbf{} LaTeX c
code. This has now been fixed.

> Two paragraphs start with "Finally" in the "Data-aware Methods" section.

Revisions: Fixed.

> Openness/Transparency
>
> As a condition for acceptance, I hope the authors can provide the stimuli, data analy
> any other applicable materials per the Jovi requirements https://www.journalovi.org
> guide.html#transparency-requirements.

Revisions: Done; we have added all of the available data we have to the OSF repository.

> The CSV file for the experimental data matches the number of participants, condition
> repeated trials described in the paper. The README describes the data columns well

the analysis scripts in the OSF link.

Excellent.

> Suggested outcome Major revisions: this paper requires substantial improvements tl
> to re-review to decide whether or not to endorse it.
>
> Requested changes Refine the knowledge contribution in the context of ATOM and p
> newer literature Clarify the bin size algorithm and justify the packing layout decisio
> experimental stimuli and data analysis scripts Resolve analysis issues and elaborate
> analysis results translate to the practical use of gatherplots.

We believe this has all been done. Thank you for the careful feedback.

@facet-fan on May 14, 2024 20:54:

Minor revisions: This paper requires some smaller changes, after which I am confident
endorse it.

Summary

I appreciate the authors addressing my comments in great detail. The authors have impr
works section, added more details and justifications on the design of Gatherplot, and refi
along with more open materials. There are a few more things the authors can address fo

Results section

1. Can the authors double-check the completion times reported in the text:

   > For the retrieve-value task (T1), on average, the completion time (sec) for each interf
   > jitter 31.79, absolute 36.83, normalized 36.55, and both 49.21.

By referencing Figure 13, I find that these values look like the lower bounds of the CI, not
this issue does not affect the conclusions.

Misc typos and issues

- The Discussion section still states that "the gatherplots technique enable**[s]*** ***people
  distribution more *quickly and more correctly", while similar language about "mo
  edited out in the revision.
- Figure 5 `x=Name` issue not fixed
- "It is worth comparing gatherplots to ***ur*** own prior work" (Section 2)
- "We chose to include an interactive setting—``both''—in our experiment" (latex style
  Section 6)
- `repetitions).}` in Figure 12 caption

@nickelm on Jun 08, 2024 09:11:

> **Results section**
>
> 1. Can the authors double-check the completion times reported in the text:
>
>   > For the retrieve-value task (T1), on average, the completion time (sec) for each i
>   > was for jitter 31.79, absolute 36.83, normalized 36.55, and both 49.21.
>
> By referencing Figure 13, I find that these values look like the lower bounds of the CI
> means, though this issue does not affect the conclusions.

Revision: Thanks for spotting this, it was a clear mistake. We have updated the completio
the text.

> **Misc typos and issues**

- The Discussion section still states that "the gatherplots technique enable**[s]*** ***pe***
  data distribution more _***quickly***_ and more correctly", while similar languag
  quickly" has been edited out in the revision.

Revision: Fixed, thank you.

- Figure 5 `x=Name` issue not fixed

Yes, we are still working on this; I'll address it during this second-round revision process.

- "It is worth comparing gatherplots to ***ur*** own prior work" (Section 2)
- "We chose to include an interactive setting—``both''—in our experiment" (latex
  marks, Section 6)
- `repetitions).}` in Figure 12 caption

Revision: All fixed, thanks!

@codementum on Jun 14, 2024 00:28:	[referenced from #[DECISION] Gatherplots 13-May-2024]
