Issue #7 (open): [REVIEW] Gatherplot review

[opened]

Conflicts of interest

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

Review

This paper seeks to make constructing scatterplots less prone to overlap when using categorical axes. To do so, they introduce a method called gatherplot/gather transform. It involves a mechanism for two-dimensional partitioning and a series of layout algorithms for presenting unit visualizations within those partitions. The algorithms are three: absolute (which is a waffle plot-like layout), normalized (space-filling waffle plot), and stream graph (a specialization of absolute, it seems). This is evaluated via a series of examples (in figures) and a crowd-work study. Various design considerations are described, including combining the gatherplot with a magic lens style interaction (gather lens).

This is generally interesting work that seeks to solve a realistic problem; however, I think it is *currently* a ways from being ready for publication. To wit, despite the clear enunciation of contributions provided in the introduction, I suggest that it is slightly ambiguous what is contributed in this work. It is suggested that there are four contributions:

1: gather visual transform
2: gather plot chart type
3: gather lens 4: the crowd work evaluation

- I would argue that (1) is not a contribution made in this work, as it is simply the idea of data-based partitions (which is well known, as this is the idea of a data cube). In addition, as discussed below, some issues with that section's presentation impede its clarity.
- While the inclusion of a magic lens is pretty interesting, I do not think it is a contribution in its current form. Why is this interaction particularly well matched with this chart type? The figure demonstrating this feature seems to show a scatterplot with substantial overlap, which would seem to suggest that it is not a gatherplot (but it is unclear). If these are not to be used in conjunction with gatherplots, then it would seem that this is a separate contribution and should be explored in its own context rather than taking the spotlight off gatherplots. Suggesting then that (3) is also yet to be a contribution.
- I think the contribution of a new plot type is certainly interesting (2), and while gatherplot resembles many other chart types (for instance, bee swarm/swarm plots), they are unusual in their modularity. However, the particular evaluation strategy does not demonstrate this is a good chart relative to other chart types. It only shows that one mode (jitter) is less accurate than others, not that it is better than comparable chart types, nor does it enumerate what it can and can not do (see below for more about this). Similarly, an important idea with developing a chart type is that it makes it easy to conceptualize things for their users—things that might be difficult to do otherwise. It would be exciting to compare gatherplots as chart type against atom as a unit grammar.
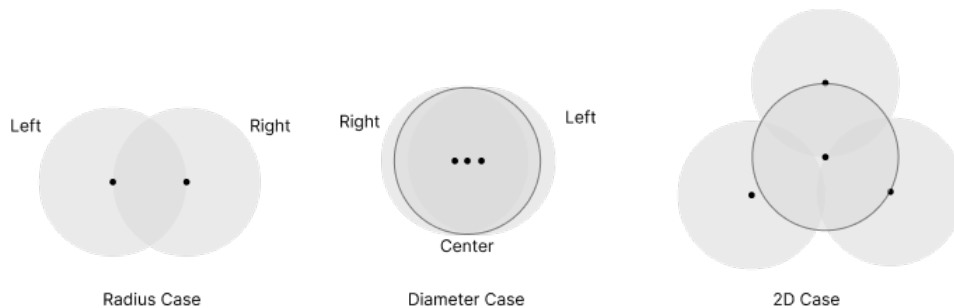
- It is also possible that the sub-algorithms (norm/abs/streamgraph) might be claimed as contributions. These are interesting algorithms, and the crowd-work study does approach evaluating them. However, there is also a very large space of additional algorithms that might be slotted into the gather transformation. For instance, in #4 @RaphaelWimmer asks about the relationship between this work and the somewhat well-known swarm plots. There are many forms of swarm plot, but a prominent one involves using simple force-direction as means to create non-overlapping layouts. Other charts could be developed as well: for instance, nothing is stopping the partitions from containing things like unit-based pies or spirals and so on. It is not demonstrated that these are the best layout algorithms for this context or even really described what the space of possible charts might be (which would be an interesting contribution, although likely quite similar to atom).

I think the main way in which InfoVis can build on this work is based on the result found in the user study that the normed/abs layouts are easier to use than just jitter—suggest perhaps a guideline that layouts other than jitter should be used, (which is a useful guideline that can be easily and directly applied elsewhere). However, the main body of the contribution (gatherplots) is somewhat under-evaluated. In addition, many components of the work that contribute to the overall construction need more care.

Evaluation: I have some additional comments/questions about the evaluation.

- what are the limitations of this approach? What can it do? What can it not do? In the current presentation, this isn't really reflected on (nor are the results of the study synthesized)
- Given that this work seeks to be in dialog with scatterplots, I was surprised not to see usage of Sarikaya and Gleicher's "Scatterplots: Tasks, Data, and Design" task typology. Several tasks are described throughout (such as highlighting outliers), but these are not drawn from a single systemic task model. I would like to see a clear description of what tasks gatherplots (ideally in each mode) succeed and fail under such designs. This might more clearly allow the reader to understand the contribution of gatherplots.
- I like that the presentation of the evaluation is fairly systematic. Could some of these points be combined in such a way that we get all of this information from a single chart? Similarly, I found it a little difficult to compare the algorithms in Figure 12, it'd be nice just to make 1 line per task (ideally with some context so we don't have to look up what the tasks are) and then pivot the algorithm by color and a small dy offset
- How much were participants paid?
- Was the stream plot tested? If not, why not?
- I would like to see an example of the stimulus shown to participants in the paper. There's no space limit here! Please be generous with visual evidence.
- I would like to see a direct comparison with/commentary on the author's own work atom. How does this compare to that? Can any of the algorithms in gatherplot be expressed as atom charts? Or does atom already do all of what this does? (Similarly there a large number of related charts mentioned in the related work, itd be useful to see a gallery of each of them trying to produce the same chart, to see how this form relates to those)
- I am not evaluating the statistics applied in the evaluation; I leave that to a reviewer with greater expertise in that style of analysis.

Mathematical Rigor: Section 3 attempts rigorously define the problem setting and proposed chart. I emphasize that I really like that the paper tries to be rigorous in its description, however, the level of rigor is far from sufficient and the resulting formalism is not used in the execution of the chart Majorly: the specific of the overlap definition is imprecise and allows for a wide variety of overlaps. For instance, it is not clear if mark size s refers to radius or diameter, each of which fails in various basic cases:

Radius Case   Diameter Case   2D Case

The radius case fails the test but is still clearly visible and should not be counted as overlapping as it can clearly be read, even at full opacity. Diameter case passes but clearly should fail as the center circle is totally overlapped. To make matters slightly more difficult, this definition is only applied *per dimension*. Overlaps in multiple dimensions may be more quarrelsome to catch points hidden by multi-dimensional overlaps (as above). In all, this makes this definition/test very, very brittle.

More minorly: many symbols in use are not defined or are hand-waved through. C is not defined, p_x is suggested as being a point, but this is not clearly defined, and D is not rigorously defined (for instance, what is its relationship with the rest of the data set?). The Describe corollary is not a corollary (in particular bc there have not been proofs), but also because it is a direct consequence of the definition used.

Given that this section has little consequence for the rest of the work, I recommend that it be removed. However, I would, of course, be happy to be wrong about the rigor of this measurement. This might be demonstrated empirically by measuring synthetic data sets with known amounts of overlap. (Similarly, I would like to see a visual argument of the "using the gather transformation" section and the "segments as labeled intervals instead of labeled major and minor ticks" in the preceding section. Claims are being made here, there should be evidence provided.)

Related

- How does this definition handle varying-sized marks (can unit visualizations not encode size as one of their attributes? Perhaps "Relaxed Dot Plots" might make a point of comparison here?)
- How does this handle not symmetric elements?

Misc

- I like that there is a certain amount of reasoning about how to effectively design with this chart type, but I wish that I found the arguments made in the visual design and interaction sections were more carefully argued. For instance, it suggested that stroke lines should be avoided all of the time, but this would seem to limit many designs needlessly. Why impose such a strict rule if it isn't necessary? It is noted that several alternate designs are considered for what mark to use, however the specifics of these design explorations are not discussed. Reciprocally, there is discussion of the interval tick marks, however the conclusion that 9f is the optimal design seems shallowly argued based on a notion of minimalism (which is itself debatable compared to 9e).
- The axis folding interaction seems quite interesting, but it isn't evaluated (also, as a viewer, I expected it to fold to an aggregate? is it not creating visual distortion in its current form?)
- I think it's part of the work's conceit that unit visualizations are worth considering, which I am willing to accept as a premise. However, the paper repeatedly dismisses aggregate marks as not being valuable, even though they are still plenty usable and unambiguous (such as in cluster views and so on). My favorite example of this usability is the Craigslist apartment finder, which uses clustering to great effect (although there are of course countless other examples). I would like to see a more careful argument why we should dismiss these forms

out of hand.

- A number of chart forms are dismissed in the intro/related work sections as they are unable to handle larger data, which is a flaw of this approach as well (for instance how would gatherplots deal with 1 million data points). Similarly, it is noted but not argued that changing "mark size, increasing display space, or animating the data" are not practical changes. It would be good to make these statements more careful.
- It looks like this paper may be a resubmission from an earlier time. To be clear: there is nothing at all wrong with resubmitting something later. However, it also appears that the paper has not been updated in at least five years. While it is true that some things are timeless, part of rigorous work is to be in dialog with contemporary scholarship and relating the value of the work to those that came before it (or after it, as the case may be)—for instance, missing Sarikaya's scatterplot task model and all of the thoughtful empirical work that appears after it (or swarm plots!) seems like an oversight. Per the reviewer guidelines, I don't hold the authors to this point but merely note it.
- There's a handful of typos throughout, some more confusing ("Plotting the gives rise to diagonal line"), some less ("$p<0.01>$", "as (or) in"), but all should be fixed
- It would be nice if there was a clear way for others to use this system outside of the provided interfaces. Consider extracting the layout algorithms somewhere reusable (like an npm package). I suggest that this type of engineering consideration helps supports the goals of openness held by this journal (what's more open then being able to try something yourself?), and so should be taken seriously. Related: it is noted in the preface that the system is implemented on observable. While I can see that this is true in the code for the article, providing a link to the observable notebook might be clearer, which I would imagine would be a more common way to interpret that assertion.
- While I am intrigued by its inclusion, I don't quite understand why Firebase is a necessary component of this system. What properties of Firebase's real-time database are utilized? Is it just for hosting? If the latter, then it does not need to be listed as part of the implementation strategy.
- Figure 5, in its pre-interaction state, may show an implementation bug (see below). It looks like it is saying that the vw rabbit custom is the whole axis when in fact, it's the name?

Figure 5: **Interactive gatherplot.** Gatherplot of the cars dataset with controls for the horizontal and vertical axes as well as the dot color assignment.

Openness/Transparency

Required - An example of the stimuli used the crowd work study - The scripts used to produce the analysis

Nice to have - a way for others to use this approach (such as an npm library or observable chart type)

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-review to decide whether or not to endorse it.

Requested changes

Major

- Remove (or make rigorous) the overlap formalism
- Remove (or more carefully argue) the design recommendations
- Remove the discussion of gather lens, or, provide convincing evidence why this is unique to approach
- Provide clearer comparison with other chart types or (preferably) clearly demonstrate what tasks can and can not be done with this chart type

Minor

- typos, presentational issues noted throughout

ORCID

No response

@codementum on
Dec 20, 2023 20:33:

[referenced from #[DECISION] Gatherplots 20-Nov-2023]

@nickelm on
Mar 24, 2024 09:52:

We thank the reviewer for this feedback. Below I have quoted the relevant parts of the review along with our responses.

> It is suggested that there are four contributions:
>
> 1: gather visual transform 2: gather plot chart type 3: gather lens 4: the crowd work evaluation
>
> I would argue that (1) is not a contribution made in this work, as it is simply the idea of data-based partitions (which is well known, as this is the idea of a data cube). In addition, as discussed below, some issues with that section's presentation impede its clarity.

Revision: We have removed the claim that the gather transformation is a contribution (however, we have chosen to revise rather than remove the section).

> While the inclusion of a magic lens is pretty interesting, I do not think it is a contribution in its current form. Why is this interaction particularly well matched with this chart type? The figure demonstrating this feature seems to show a scatterplot with substantial overlap, which would seem to suggest that it is not a gatherplot (but it is unclear). If these are not to be used in conjunction with gatherplots, then it would seem that this is a separate contribution and should be explored in its own context rather than taking the spotlight off gatherplots. Suggesting then that (3) is also yet to be a contribution.

This is a fair point. We felt that the GatherLens serves as an interesting extension to the gatherplot idea. In fact, we employed a variant of it in our own prior work on interactive topic modeling, where we found it an effective interaction. However, we

hear this feedback clearly. Furthermore, the GatherLens implementation is no longer maintained and was not replicated in our clean-room implementation of gatherplots.

Revision: We have removed Section 6 from the paper as well as the third contribution.

> I think the contribution of a new plot type is certainly interesting (2), and while gatherplot resembles many other chart types (for instance, bee swarm/ swarm plots), they are unusual in their modularity. However, the particular evaluation strategy does not demonstrate this is a good chart relative to other chart types. It only shows that one mode (jitter) is less accurate than others, not that it is better than comparable chart types, nor does it enumerate what it can and can not do (see below for more about this). Similarly, an important idea with developing a chart type is that it makes it easy to conceptualize things for their users—things that might be difficult to do otherwise. It would be exciting to compare gatherplots as chart type against atom as a unit grammar.

We agree. Our primary goal at the time of the evaluation (Spring 2014) was to compare the approach to jittering, which we felt was a reasonable baseline. Knowing what we know today, we would likely have included a wider set of comparisons.

Unfortunately, with the original students graduated and in their own faculty positions, we are not in a position to extend the study.

> It is also possible that the sub-algorithms (norm/abs/streamgraph) might be claimed as contributions. These are interesting algorithms, and the crowd-work study does approach evaluating them. However, there is also a very large space of additional algorithms that might be slotted into the gather transformation. For instance, in Difference to Swarmplots? [#4](#) @RaphaelWimmer asks about the relationship between this work and the somewhat well-known swarm plots. There are many forms of swarm plot, but a prominent one involves using simple force-direction as means to create non-overlapping layouts. Other charts could be developed as well: for instance, nothing is stopping the partitions from containing things like unit-based pies or spirals and so on. It is not demonstrated that these are the best layout algorithms for this context or even really described what the space of possible charts might be (which would be an interesting contribution, although likely quite similar to atom).

All good points. We had already discussed using glyphs and aggregated visual representations in the partitions in the "Visual Marks" section. We have now revised and improved this section to better clarify our focus on unit marks in the paper, but noting that aggregate marks are also possible.

As to the comment about swarmplots, this is fair. We have responded to @RaphaelWimmer's issue as well.

Revision: We have added a discussion of stripcharts, beeswarm plots, swarmplots, and stripplots to the subsection "Visualizing Categorical Variables".

> I think the main way in which InfoVis can build on this work is based on the result found in the user study that the normed/abs layouts are easier to use than just jitter—suggest perhaps a guideline that layouts other than jitter should be used, (which is a useful guideline that can be easily and directly applied elsewhere). However, the main body of the contribution (gatherplots) is somewhat under-evaluated. In addition, many components of the work that contribute to the overall construction need more care.

Thank you. We have tried to make improvements throughout the paper based on this review.

> Evaluation: I have some additional comments/questions about the

evaluation.

> what are the limitations of this approach? What can it do? What can it not do? In the current presentation, this isn't really reflected on (nor are the results of the study synthesized)

Revision: This is a good point. We have added a new Discussion section to the paper where we, among other things, discuss the limitations of the technique.

> Given that this work seeks to be in dialog with scatterplots, I was surprised not to see usage of Sarikaya and Gleicher's "Scatterplots: Tasks, Data, and Design" task typology. Several tasks are described throughout (such as highlighting outliers), but these are not drawn from a single systemic task model. I would like to see a clear description of what tasks gatherplots (ideally in each mode) succeed and fail under such designs. This might more clearly allow the reader to understand the contribution of gatherplots.

This is a good point.

Revision: We have added and discussed Sarikaya and Gleicher's paper, and used it to motivate our evaluation.

> I like that the presentation of the evaluation is fairly systematic. Could some of these points be combined in such a way that we get all of this information from a single chart? Similarly, I found it a little difficult to compare the algorithms in Figure 12, it'd be nice just to make 1 line per task (ideally with some context so we don't have to look up what the tasks are) and then pivot the algorithm by color and a small dy offset

Revision: We have added color to the charts and rearranged them based on task. This makes it easier to compare across visualizations.

> How much were participants paid?

Participants were compensated for their effort consistent with U.S. federal minimum wage ($7.25/hour).

> Was the stream plot tested? If not, why not?

It was not tested. The stream plot is a more specialized form of the gatherplot, and we did not feel it was comparable to other layout modes for this reason.

> I would like to see an example of the stimulus shown to participants in the paper. There's no space limit here! Please be generous with visual evidence.

Revisions: We have added an example run of the evaluation to the OSF, and we have also added a single example of the visual stimulus as Figure 11 in the paper.

> I would like to see a direct comparison with/commentary on the author's own work atom. How does this compare to that? Can any of the algorithms in gatherplot be expressed as atom charts? Or does atom already do all of what this does? (Similarly there a large number of related charts mentioned in the related work, itd be useful to see a gallery of each of them trying to produce the same chart, to see how this form relates to those)

Revision: We have added a direct comparison to the end of the subsection titled "Data-aware Methods".

> I am not evaluating the statistics applied in the evaluation; I leave that to a reviewer with greater expertise in that style of analysis.

This is fair. We have uploaded our analysis script on OSF for this review.

> Mathematical Rigor: Section 3 attempts rigorously define the problem setting

and proposed chart. I emphasize that I really like that the paper tries to be rigorous in its description, however, the level of rigor is far from sufficient and the resulting formalism is not used in the execution of the chart Majorly: the specific of the overlap definition is imprecise and allows for a wide variety of overlaps. For instance, it is not clear if mark size s refers to radius or diameter, each of which fails in various basic cases:

Yes, this is absolutely a fair point of criticism. The formalism is not even necessary.

The radius case fails the test but is still clearly visible and should not be counted as overlapping as it can clearly be read, even at full opacity. Diameter case passes but clearly should fail as the center circle is totally overlapped. To make matters slightly more difficult, this definition is only applied per dimension. Overlaps in multiple dimensions may be more quarrelsome to catch points hidden by multi-dimensional overlaps (as above). In all, this makes this definition/test very, very brittle.

Agreed.

More minorly: many symbols in use are not defined or are hand-waved through. C is not defined, p_x is suggested as being a point, but this is not clearly defined, and D is not rigorously defined (for instance, what is its relationship with the rest of the data set?). The Describe corollary is not a corollary (in particular bc there have not been proofs), but also because it is a direct consequence of the definition used.

Again, we agree.

Given that this section has little consequence for the rest of the work, I recommend that it be removed. However, I would, of course, be happy to be wrong about the rigor of this measurement. This might be demonstrated empirically by measuring synthetic data sets with known amounts of overlap. (Similarly, I would like to see a visual argument of the "using the gather transformation" section and the "segments as labeled intervals instead of labeled major and minor ticks" in the preceding section. Claims are being made here, there should be evidence provided.)

Revision: We have followed the advice and removed the unnecessary formalism from the paper. The subsection on "Problem Definition" (defining overlapping) is now entirely removed. We have revised the subsection "Definition: Gather Transformation" to eliminate the formalism.

How does this definition handle varying-sized marks (can unit visualizations not encode size as one of their attributes? Perhaps "Relaxed Dot Plots" might make a point of comparison here?)

Varying-size marks are outside the scope of the gatherplots technique for now.

How does this handle not symmetric elements?

Again, agreed. Our more "common sense" definition describes this better.

I like that there is a certain amount of reasoning about how to effectively design with this chart type, but I wish that I found the arguments made in the visual design and interaction sections were more carefully argued. For instance, it suggested that stroke lines should be avoided all of the time, but this would seem to limit many designs needlessly. Why impose such a strict rule if it isn't necessary? It is noted that several alternate designs are considered for what mark to use, however the specifics of these design explorations are not discussed. Reciprocally, there is discussion of the interval tick marks, however the conclusion that 9f is the optimal design seems shallowly argued based on a notion of minimalism (which is itself

debatable compared to 9e).

These are fair points.

Revision: We have attempted to reduce these strict "rules" throughout the Gatherplots section.

> The axis folding interaction seems quite interesting, but it isn't evaluated (also, as a viewer, I expected it to fold to an aggregate? is it not creating visual distortion in its current form?)

Yes, we did not evaluate axis folding. Folding to an aggregate would be one design alternative that we did not investigate. However, we do use a standard tick mark to communicate that the axis has been folded (Figure 10).

> I think it's part of the work's conceit that unit visualizations are worth considering, which I am willing to accept as a premise. However, the paper repeatedly dismisses aggregate marks as not being valuable, even though they are still plenty usable and unambiguous (such as in cluster views and so on). My favorite example of this usability is the Craigslist apartment finder, which uses clustering to great effect (although there are of course countless other examples). I would like to see a more careful argument why we should dismiss these forms out of hand.

This is a fair point. We were over-eager in evangelizing the merits of gatherplots in the paper.

Revisions: We have softened these statements throughout the paper.

> A number of chart forms are dismissed in the intro/related work sections as they are unable to handle larger data, which is a flaw of this approach as well (for instance how would gatherplots deal with 1 million data points). Similarly, it is noted but not argued that changing "mark size, increasing display space, or animating the data" are not practical changes. It would be good to make these statements more careful.

Another fair point; ibid above.

> It looks like this paper may be a resubmission from an earlier time. To be clear: there is nothing at all wrong with resubmitting something later. However, it also appears that the paper has not been updated in at least five years. While it is true that some things are timeless, part of rigorous work is to be in dialog with contemporary scholarship and relating the value of the work to those that came before it (or after it, as the case may be)—for instance, missing Sarikaya's scatterplot task model and all of the thoughtful empirical work that appears after it (or swarm plots!) seems like an oversight. Per the reviewer guidelines, I don't hold the authors to this point but merely note it.

Yes, this is well-spotted. The paper seemed a good fit for the JoVI experimental interactive article track, so here we are.

> There's a handful of typos throughout, some more confusing ("Plotting the gives rise to diagonal line"), some less ("$p<0.01>$", "as (or) in"), but all should be fixed

Revision: Thanks. We have attempted to fix these problems.

> It would be nice if there was a clear way for others to use this system outside of the provided interfaces. Consider extracting the layout algorithms somewhere reusable (like an npm package). I suggest that this type of engineering consideration helps supports the goals of openness held by this journal (what's more open then being able to try something yourself?), and

so should be taken seriously. Related: it is noted in the preface that the system is implemented on observable. While I can see that this is true in the code for the article, providing a link to the observable notebook might be clearer, which I would imagine would be a more common way to interpret that assertion.

We have followed the JoVI interactive article method of embedding the code in the article itself. The Quarto format internally uses Observable notebooks for its interactive features. In other words, the article *is* the Observable implementation.

> While I am intrigued by its inclusion, I don't quite understand why Firebase is a necessary component of this system. What properties of Firebase's real-time database are utilized? Is it just for hosting? If the latter, then it does not need to be listed as part of the implementation strategy.

This was merely for hosting purposes.

Revision: We have removed Firebase as part of the implementation strategy.

> Figure 5, in its pre-interaction state, may show an implementation bug (see below). It looks like it is saying that the vw rabbit custom is the whole axis when in fact, it's the name?

Yes, this is a bug. We are addressing it.

> Suggested outcome Major revisions: this paper requires substantial improvements that I will need to re-review to decide whether or not to endorse it.

We have revised the paper based on this review and hope that the new version will meet with the reviewer's approval. Thanks again for the careful attention to detail!

@joviewer-xyz on May 02, 2024 22:01:

I thank the authors for their hard work in updating the paper. It is substantially improved from the previous version. I appreciate the streamlined focus of the paper, which now more clearly enunciates its contribution as being centered on the gather plot chart type and corresponding user study. Further it much more clearly positions itself within work that has occurred after its initial submission.

I read the paper as now offering a solution to the problems of over-plotting in scatterplot and the data distortion caused by a common approach to resolving that issue while maintaining unit visualizations. Within that tighter scope the evaluation does demonstrate that this strategy succeeds in that goal, for a particular range and set of parameters. I suggest that the last piece needed to finish this work (beside some additional textual clean up) would be to characterize the parameter ranges in which this should be used. How many points, how many categorical variables can be used, and so on. To this end: an argument made early on is that over-plotting does not scale for larger datasets, however it seems likely that gather plot also does not scale to larger datasets. I appreciated the added note in the discussion to this effect, but this work would be made much stronger if ballpark estimates on when it would be appropriate to use a different chart type. Ideally such questions could be answered through another crowd work study. However, @nickelm has emphasized in responding to all three reviews that additions to the study are not possible because the student has departed, suggesting that such requests are unlikely to be met (although, I will admit, this is a frustrating response given the apparent simplicity of the study conducted). In light of this I suggest that a simulation study, such as exploring the differentiability of points in pixel space (or applying the definitions of legibility/accuracy in a computational manner) might also address these concerns and could be done with only the implementation actively being used on the web page.

I have a few other notes that should be addressed in revision, but I think this work is generally on a good course.

Major

- It is repeatedly emphasized (through the intro/related works) that aggregated plots are not sufficient because they forgo the unit visibility. While this is essentially an axiom for unit visualizations, I am not sure that it is convincingly argued: why is it important that these values be shown? If there are so many points that aggregation is possibly necessary/appropriate then is the value of unit-izing? For instance, I suggest that in the user study something like a pivot table would probably do just as good (if not better) than all of the stimuli examined. I implied this in my first review and make explicit it here, because an important part of characterizing the space of useful applications requested should also highlight when actually using gather plots will be useful.

- I found the "using the gather transformation" subsection to be incomprehensible. The first paragraph feels like there may have been some splicing with previous arguments that have left it to be pretty unclear. The second paragraph seems to be arguing that the axes are a visualization, which I find to be generally unconvincing in its current form. The figure just below it does not seem to support this idea either. Comparison with parallel sets is an interesting idea, but if this is an important contribution (and from my reading of the other reviews/responses it is not) then this section should be re-written and this argument made more explicit. Alternative (and preferably) this section should be removed.

- There are still some lines of reasoning about the design guidelines (cf "we recommend using a rectangle with constant rounded edge.", "on the order of 10 or less") that are not clearly argued. I continue to recommend that these be removed or substantially more care put into making those arguments. For instance, to the former of these referenced arguments, this could be done with a figure. However, given my basic request of "better characterizing the design and parameter space", it would be good to not remove this argument, but expand it: there are many other mark types that are possible (eg text, emoji, etc) it would be good to characterize the design space more clearly. Related: Perhaps instead it would be good to emphasize (at least in the "Visual Design" section) that these choices are not specifically necessary to the contribution but are merely the ones explored in this implementation. Can more commentary/reflection be offered in this section to guide future usage?

Minor

- The definition of legibility / accuracy seems to be VERY tightly bound to a notion of unit visualizations. it would be good to justify these slightly unusual uses of these terms.

- I would think that a more conventional response to seeing a highly over-plotted scatterplot like the one in Figure 7 would be to facet it by color so that there was less visual conflict? This would be a good case to add. Per my earlier request for a more formal accounting of the limits of this chart form, it would be useful to make the number of points in the chart resizable (e.g. a slider for N) so that the reader might be able to see for themselves how well this does at different granularities.

- my comment about alternate layouts within cells is not addressed in the update, I still am curious about the potential of other spatial arrangements (such as pie charts and so on). In the response it is suggested that the Visual Marks subsection addresses this concern, but I emphasize that my question was about the arrangement of the units within an aggregation cell and not the content of the marks.

- It would be interesting to explore the design space of marginalia for this chart

type. For instance, in scatterplots strip plots or kdes are sometimes used to show a 1d depiction of a distribution along one axis. Related: How do grid lines fit into the design space of this chart type?

- "same continuous variable on both axes" gives rise to an interesting 1D visualization. It might be worth comparing with Correll's studies on that genre ("Teru Teru Bōzu: Defensive Raincloud Plots" and "Looks good to me: Visualizations as sanity checks")

- Presentation notes: It might be clearer to merge the gather transform and gather plot sections as they are pretty deeply intertwined. Contrastingly, I found the presentation of the gather transform to be pretty fast, maybe my reading is biased by the previous version, but I was expecting a slower and more explicit description of the transform (perhaps via a diagram)? Some places to be more precise:

  "simplest among them are histograms"

Is this the correct name for this chart type? I would have thought that bar chart is meant here, as histograms typically convey binning of a quantitative variable, but perhaps there is some history I am not aware of.

  Wilkinson proposed .25n^-1/2 as the optimal dot size for dot plots

What is n in this case

  It also introduces uncertainty that is not aptly communicated by the scatter plot since marks will no longer be placed at their true location on the Cartesian space.

I believe that this is not a conventional usage of the term uncertainty. I believe distortion is intended.

  Some trivial but impractical

Still don't buy that these are trivial or impractical. The following argument does not justify this harsh position.

- There are still a number of typos throughout. For instance, some inaccuracies with made cuts, for instance gather lens/parallel sets is mentioned in the intro and conclusion, there are still some stray $s, "to ur own"

Round 2 Decision

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

I thank the authors for their hard work in updating the paper. It is substantially improved from the previous version. I appreciate the streamlined focus of the paper, which now more clearly enunciates its contribution as being centered on the gather plot chart type and corresponding user study. Further it much more clearly positions itself within work that has occurred after its initial submission.

Thank you.

I read the paper as now offering a solution to the problems of over-plotting in scatterplot and the data distortion caused by a common approach to resolving that issue while maintaining unit visualizations. Within that tighter scope the evaluation does demonstrate that this strategy succeeds in that

goal, for a particular range and set of parameters. I suggest that the last piece needed to finish this work (beside some additional textual clean up) would be to characterize the parameter ranges in which this should be used. How many points, how many categorical variables can be used, and so on. To this end: an argument made early on is that over-plotting does not scale for larger datasets, however it seems likely that gather plot also does not scale to larger datasets. I appreciated the added note in the discussion to this effect, but this work would be made much stronger if ballpark estimates on when it would be appropriate to use a different chart type. Ideally such questions could be answered through another crowd work study. However, @nickelm has emphasized in responding to all three reviews that additions to the study are not possible because the student has departed, suggesting that such requests are unlikely to be met (although, I will admit, this is a frustrating response given the apparent simplicity of the study conducted). In light of this I suggest that a simulation study, such as exploring the differentiability of points in pixel space (or applying the definitions of legibility/accuracy in a computational manner) might also address these concerns and could be done with only the implementation actively being used on the web page.

I thank the reviewer for their understanding. While nominally a new study could be doable, I recently moved institution (and country), and find myself in a new environment with three-year Ph.D.s where asking a new student to take on an old project such as this seems unfair.

Revision: The idea of quantifying the scalability of gatherplots is sound. I have added a new subsection to Section 4 called "Visual and Data Scalability" where I discuss an upper bound for the number of visual marks that can be shown in a gatherplot, especially in "pathological" situations where not all the available display space is used by a visual mark. I think that this is a nice addition to the paper, and hope that this treatment will satisfy the reviewer.

I have a few other notes that should be addressed in revision, but I think this work is generally on a good course.

Major

- It is repeatedly emphasized (through the intro/related works) that aggregated plots are not sufficient because they forgo the unit visibility. While this is essentially an axiom for unit visualizations, I am not sure that it is convincingly argued: why is it important that these values be shown? If there are so many points that aggregation is possibly necessary/appropriate then is the value of unit-izing? For instance, I suggest that in the user study something like a pivot table would probably do just as good (if not better) than all of the stimuli examined. I implied this in my first review and make explicit it here, because an important part of characterizing the space of useful applications requested should also highlight when actually using gather plots will be useful.

Revision: Thanks for this feedback. I have added a new subsection called "Visibility, Discriminability, and Spatial Accuracy" that discusses this explicitly. This section also brings in the two metrics "spatial accuracy" and "discriminability" (changed from "legibility"; see below) to a more logical place in the article (they were previously introduced in passing in the "Managing Continuous Variables" subsection).

- I found the "using the gather transformation" subsection to be incomprehensible. The first paragraph feels like there may have been some splicing with previous arguments that have left it to be pretty unclear. The second paragraph seems to be arguing that the axes are a visualization, which I find to be generally unconvincing in its current

form. The figure just below it does not seem to support this idea either. Comparison with parallel sets is an interesting idea, but if this is an important contribution (and from my reading of the other reviews/ responses it is not) then this section should be re-written and this argument made more explicit. Alternative (and preferably) this section should be removed.

Revision: Thank you for highlighting this. I agree and have removed the section. I am not sure where this wording came from originally, but it was certainly orphaned in the current version of the paper.

- There are still some lines of reasoning about the design guidelines (cf "we recommend using a rectangle with constant rounded edge.", "on the order of 10 or less") that are not clearly argued. I continue to recommend that these be removed or substantially more care put into making those arguments. For instance, to the former of these referenced arguments, this could be done with a figure. However, given my basic request of "better characterizing the design and parameter space", it would be good to not remove this argument, but expand it: there are many other mark types that are possible (eg text, emoji, etc) it would be good to characterize the design space more clearly. Related: Perhaps instead it would be good to emphasize (at least in the "Visual Design" section) that these choices are not specifically necessary to the contribution but are merely the ones explored in this implementation. Can more commentary/reflection be offered in this section to guide future usage?

Revision: We have removed these motivations; for one thing, they were not consistent now that we have two implementations with slightly different designs.

Minor

- The definition of legibility / accuracy seems to be VERY tightly bound to a notion of unit visualizations. it would be good to justify these slightly unusual uses of these terms.

Revision: Good point. After some reading, I decided to replace "legibility" with "discriminability" and add appropriate references to prior art using this term. I have also expanded the discussion in the subsection "Managing Continuous Variables" somewhat.

- I would think that a more conventional response to seeing a highly over-plotted scatterplot like the one in Figure 7 would be to facet it by color so that there was less visual conflict? This would be a good case to add. Per my earlier request for a more formal accounting of the limits of this chart form, it would be useful to make the number of points in the chart resizable (e.g. a slider for N) so that the reader might be able to see for themselves how well this does at different granularities.

This is a good idea, but Figure 7 is not one of the interactive charts so I have not added this feature.

Revision: I have added a note that there are alternatives based on faceting by interaction or by animation to the end of the paragraph discussing Figure 7.

- my comment about alternate layouts within cells is not addressed in the update, I still am curious about the potential of other spatial arrangements (such as pie charts and so on). In the response it is suggested that the Visual Marks subsection addresses this concern, but I emphasize that my question was about the arrangement of the units within an aggregation cell and not the content of the marks.

Thanks for clarifying.

Revision: I have added a subsubsection on "Stacked Group Layout" to the VIsual Design subsection that discusses different spatial arrangements inside the cells. I also created a figure to illustrate. I hope this is closer to what the reviewer had in mind.

- It would be interesting to explore the design space of marginalia for this chart type. For instance, in scatterplots strip plots or kdes are sometimes used to show a 1d depiction of a distribution along one axis. Related: How do grid lines fit into the design space of this chart type?

Revision: We hope that the above new subsubsection will serve. Furthermore, I added a new subsubsection that briefly discusses the role of grid lines in a gatherplot.

- "same continuous variable on both axes" gives rise to an interesting 1D visualization. It might be worth comparing with Correll's studies on that genre ("Teru Teru Bōzu: Defensive Raincloud Plots" and "Looks good to me: Visualizations as sanity checks")

Revision: Thanks for the observation. We have discussed this case in particular in the subsection called "Managing Continuous Variables" and added citations to the two above papers.

- Presentation notes: It might be clearer to merge the gather transform and gather plot sections as they are pretty deeply intertwined. Contrastingly, I found the presentation of the gather transform to be pretty fast, maybe my reading is biased by the previous version, but I was expecting a slower and more explicit description of the transform (perhaps via a diagram)? Some places to be more precise:

I think there is value to keeping the transform and the plot separate, so I did not end up following this feedback. However, I hope that the new revised Gatherplot Transformation section (see above for my revisions to this section) will better serve.

"simplest among them are histograms"

Is this the correct name for this chart type? I would have thought that bar chart is meant here, as histograms typically convey binning of a quantitative variable, but perhaps there is some history I am not aware of.

Revision: Good point; I have revised this to "bar chart histogram."

Wilkinson proposed .25n^-1/2 as the optimal dot size for dot plots

What is n in this case

"n" is the number of data points.

Revision: I have updated the text to include this information as well as add a citation to Wilkinson (1999).

It also introduces uncertainty that is not aptly communicated by the scatter plot since marks will no longer be placed at their true location on the Cartesian space.

I believe that this is not a conventional usage of the term uncertainty. I believe distortion is intended.

Revision: Fixed, thank you!

Some trivial but impractical

Still don't buy that these are trivial or impractical. The following argument

does not justify this harsh position.

Revision: Good point. We have revised this unnecessarily harsh statement.

- There are still a number of typos throughout. For instance, some inaccuracies with made cuts, for instance gather lens/parallel sets is mentioned in the intro and conclusion, there are still some stray $s, "to ur own"

Revision: Thank you; I have tried to fix these throughout the paper.

[referenced from #[DECISION] Gatherplots 13-May-2024]