Visualizing Statistical Mix Effects and Simpson's Paradox

Zan Armstrong and Martin Wattenberg

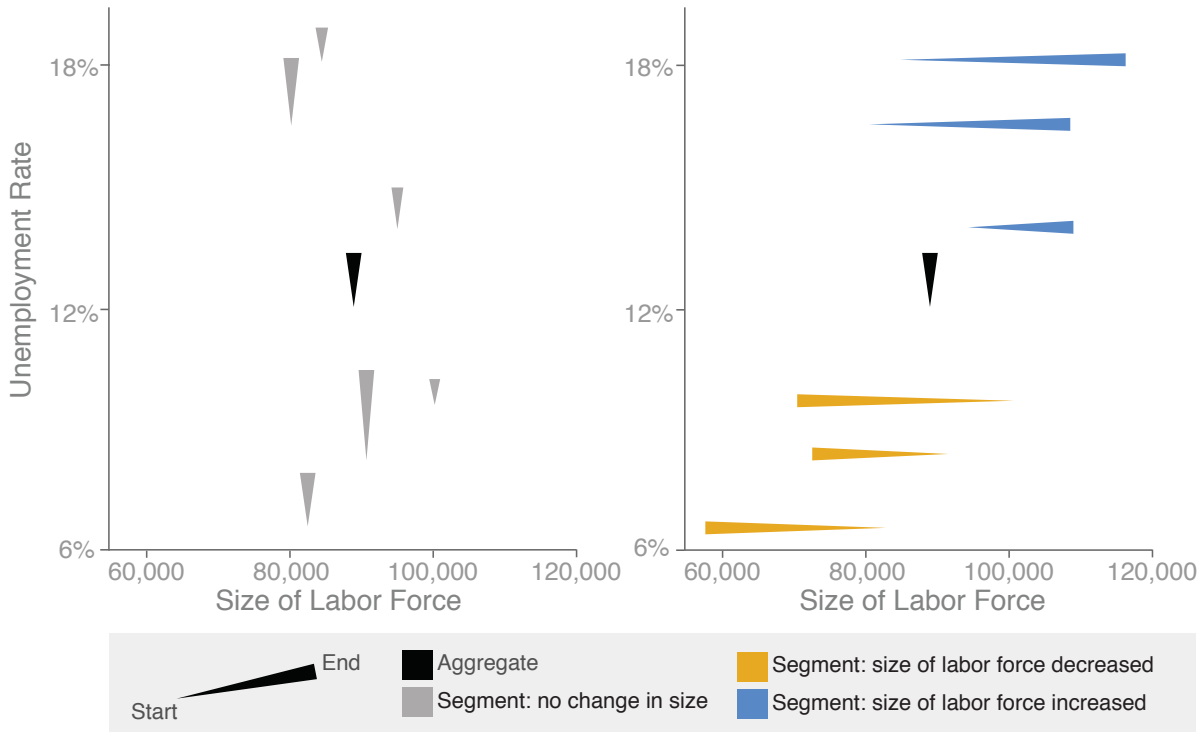

Fig. 1: Two comet charts. Left: Aggregate reflects individual segments. Right: Aggregate affected by mix effects.

Abstract—We discuss how "mix effects" can surprise users of visualizations and potentially lead them to incorrect conclusions. This statistical issue (also known as "omitted variable bias" or, in extreme cases, as "Simpson's paradox") is widespread and can affect any visualization in which the quantity of interest is an aggregated value such as a weighted sum or average. Our first contribution is to document how mix effects can be a serious issue for visualizations, and we analyze how mix effects can cause problems in a variety of popular visualization techniques, from bar charts to treemaps. Our second contribution is a new technique, the "comet chart," that is meant to ameliorate some of these issues.

Index Terms—Mix effects, Omitted variable bias, Simpson's paradox, Statistics

1 Introduction

Imagine you're an economist analyzing US unemployment. Looking at data from 2000 to 2013, you see that overall median weekly wages went up by 0.9%, after adjusting for inflation [2, 4]. A natural question is how different population segments fared. For instance, if you group the data by level of education, you'll presumably see some groups doing better than 0.9% and some doing worse. Since this has obvious policy implications, you reanalyze the data and produce Table 1.

To your surprise, *all* segments lost ground, even though the trend overall was upward! Is that really possible? Could there be something wrong with the data?

In fact, the data is correct, and what you're seeing is known to statisticians as "Simpson's paradox" [29, 34]. The resolution of this apparent paradox can be seen in Table 2.

Table 1: Change in Median Wage by Education from 2000 to 2013

Segment	Change in Median Wage (%)
Overall	+0.9%
No degree	-7.9%
HS, no college	-4.7%
Some college	-7.6%
Bachelor's +	-1.2%

Although median wages went down in each segment, something else happened as well: the number of jobs increased for higher educated groups and declined for lower. Thus the higher-educated, and therefore higher-earning, group had more weight in the 2013 data. The summary number of +0.9% depends not just on the change within population segments, but on the change in the relative sizes of those segments. The counterintuitive fact that an aggregate measure can contra-

- *Zan Armstrong was with Google at the time of research, currently unaffiliated. E-mail: zan.armstrong@gmail.com.*
- *Martin Wattenberg is with Google. E-mail: wattenberg@google.com.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346297

Table 2: Number Employed (in millions) by Education: 2000, 2013

Segment	Employed 2000	Employed 2013	Change (%)
Overall	89.4	95.0	+6.4%
No degree	8.8	7.0	-21.3%
HS, no college	28.0	25.0	-10.6%
Some college	24.7	26.0	+5.4%
Bachelor's +	27.8	37.0	+33.0%

dict all subpopulation measures is known as Simpson's paradox.

This is a real-life example, and in fact caused a minor controversy surrounding an April 2013 article in the *New York Times* [21] which contained a description of this data accompanied by multiple graphs. Five days later, the *Times* published a follow-up article because readers had noticed the discrepancy between the direction of change in wages overall and the direction of change in wages within subpopulations.

Not only did readers react with surprise but, as the author put it, "some readers deemed it evidence that I must have made a mistake." This reaction illustrates the natural difficulty that many people have when interpreting statistics that are affected by "mix"–a difficulty made worse when data is presented using standard graphics that omit the changing sizes of subpopulations.

1.1 Simpson's Paradox and Mix Effects

Simpson's paradox is an extreme example of a more general class of phenomena informally known as "mix effects": the fact that aggregate numbers can be affected by changes in the relative *size* of the subpopulations as well as the relative values within those subpopulations. One can also see this as a case of what statisticians call "omitted variable bias" or "confounding covariates," where an unexamined dimension of the data has an effect. Social scientists refer more generally to resulting misleading inferences about parts and wholes as the "ecological fallacy."

By whatever name, however, mix effects are ubiquitous. Experienced analysts encounter them frequently, and it's easy to find examples across domains. A famous example of a Simpson's-like reversal is a Berkeley Graduate Admissions study in which 44% of males were accepted but only 35% of females. The discrepancy seemed to be clear evidence of discrimination, yet disappeared once analyzed at the per-department level: it turned out that departments with lower acceptance rates had proportionally more female applicants. Similar reversals have appeared in studies of race and the death penalty in Florida [23], standardized test scores and education spending per student [14], studies of the "hot hand" in basketball [32], baseball hitting averages [25], treatments for kidney stones [10, 19], and in an active debate around the diagnosis of meningococcal disease [15].

Mix effects that don't involve the dramatic reversal of Simpson's paradox are in some ways even more dangerous. It's easy to imagine a scenario in which, say, wages drop 1% overall, but only 0.5% in all subpopulations. In such cases an analyst should note that the 1% drop comes from a combination of factors. But, without the red flag of a differing sign, such a situation may not draw the attention it deserves.

One might hope for mathematical techniques that would resolve these issues, but unfortunately there is no statistical magic bullet. For example, adding control variables to regression analysis can potentially make estimates *less* accurate [11]. A deeper issue is that disaggregation, even on all relevant variables, does not necessarily produce the correct interpretation in theory[1] [6] or practice [24]. The "standardization" approach, which tries to simulate "apples-to-apples" populations for comparison purposes [12], is a technique commonly used to adjust statistics shown in maps in order to control for the demographics [17, 22]. Unfortunately this and related methods all have known shortcomings[2] [13].

The point is that even when one can reasonably attribute a change in value to lower-level changes in mix, that alone cannot tell the correct interpretation. Additionally, sometimes the essential thing is not to adjust for mix but to discover why mix is shifting. Therefore, it is important to be able to examine aggregate changes together with the disaggregated data.

1.2 Goal of the paper

Since there is no mechanical formula for protecting against mix effects during an analysis, visualization can potentially play a critical role in helping guide analysts toward useful questions and correct conclusions. To find and communicate the real story behind the data, it would be useful to have visualization methods that portray mix effects accurately. In fact, given the pervasive nature of mix effects, it's something of a scandal that no standard business graphics do so.

In this paper we make two contributions. First, we discuss ways that mix effects can hamper interpretation of several common visualization types. For example, in certain situations, treemaps can be completely misleading. We believe these pitfalls should be more widely known. Second, we propose a simple visualization, aimed at expert analysts, that does show both value and weight at once. Our proposed visualization is based on the needs of, and feedback from, experienced data analysts at a Fortune 500 corporation.

This paper may be viewed through the lens of a design study; to use the framing given by Sedlmair et al. [27], we begin by investigating the real-world problem of mix effects. Section 2 provides analysis of the appropriate abstractions and Section 3 provides background on the problem, or "preconditions" for the work, while Section 4 describes the design of one solution, or "inward facing validation" [27]. In subsequent sections we present results from a field trial and reflect on the implications for future research, in other words our analysis and "outward-facing validation."

2 Formalizing the Problem

Mix effects can appear whenever we compare aggregated and non-aggregated data. Mathematically, we can formalize one of the most common situations–that of weighted averages and sums–as follows.

Imagine we're interested in two quantities:

$$P = \sum_{i=1}^{n} a_i x_i$$

and

$$Q = \sum_{i=1}^{n} b_i y_i$$

Here the a_i and b_i are "weights" and the x_i and y_i are "values", corresponding to statistics for n different *subpopulations* or *segments* of an overall population. The two real-world situations producing P and Q may be referred to as *scenarios*.

In the Berkeley admissions case, assuming n departments, the weights a_i and b_i could be the numbers of women and men applying to each department, and the values x_i and y_i are the acceptance rates for women and men, respectively, in each department. The quantities P and Q are the total women and men admitted.

The special case where

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = 1$$

[1] For example, Arah warns that: "It cannot be overemphasized that although these paradoxes reveal the perils of using statistical criteria to guide causal analysis, they hold neither the explanations of the phenomenon they depict nor the pointers on how to avoid them. The explanations and solutions lie in causal reasoning which relies on background knowledge, not statistical criteria." [6]

[2] Specifically, "plotting observed rates can have serious drawbacks when sample sizes vary by area, since very high (and low) observed rates are found disproportionately in poorly-sampled areas. Unfortunately, adjusting the observed rates to account for the effects of small-sample noise can introduce an opposite effect, in which the highest adjusted rates tend to be found disproportionately in well-sampled areas." [13]

corresponds to taking a weighted average of the x_i and y_i.

In a weighted average model for the Berkeley case, one could set a_i to the fraction of all women who applied to each department and b_i to the fraction of men. The quantities P and Q would then be the university-wide acceptance rates for women and men.

Written in this form, it's clear that the two quantities of interest, P and Q, each depend on two dimensions: respectively, the individual values x_i and y_i and the "mixes," or different weights, a_i and b_i. As a result, there are four relevant dimensions to the analysis.

Common visualization methods are generally not equipped to show all four dimensions at once. For example, a bar chart might compare P and Q directly, and then allow a drill-down showing comparison bar charts of only the x_i and y_i. Without showing the weights a_i and b_i as well, the resulting picture is obviously incomplete.

2.1 Absolute and relative numbers both matter

In practice, the situation can feel even more complex. The ratios between weights and values in the two scenarios–that is, a_i/b_i and x_i/y_i–are sometimes more important than the absolute numbers. For instance, in the Berkeley example, a key metric was the ratio of female to male applicants per department, a_i/b_i.

A priori, we don't know whether absolute or relative numbers will be most helpful in an analysis. There could actually be interactions between any of 6 quantities: values x_i, y_i, weights a_i, b_i, and relative ratios x_i / y_i and a_i / b_i. As a result, a tool designed for general exploration should provide a way to look at all six quantities at once to discover which combinations matter.

2.2 More general aggregation

The same issues can arise with forms of aggregation other than averages or sums. We have already seen one example: medians for unemployment data. Like weighted averages, the population median is generally not equal to the median of the subpopulation medians.

3 How mix effects can mislead in visualizations

Although the problem with omitting weights in a visualization may be obvious in the abstract, it's rarely apparent in practice. The reaction to the *New York Times* article in the introduction is an excellent example.

Even for viewers who understand the subtleties of mix effects, the issue remains that standard business graphics don't provide ways to show both value and weight comparisons at once. This makes it very difficult to tease apart the effect of mix from the effect of changes in value or to notice relationships between value and changes in size.

Mix effects can cause particular problems for interactive visualizations: as interactivity becomes more common, slicing and dicing data into segments becomes a routine action. In a world where static graphs are painstakingly prepared for publication by professional statisticians, one might hope that any chart would provide proper context. Today, however, lay users and experts alike can flip quickly through multiple ways of segmenting data sets. As the reaction to the New York Times article shows, even among a relatively sophisticated audience (readers of the *NYT* business section) a significant number of people don't realize they need to account for mix effects.

3.1 Mix effects in standard business graphics

The essential problem with standard business graphics (bar, line and pie charts) is that they show only a subset of the relevant six dimensions, often values only (as in a bar or line chart) or weights only (as in a pie chart). Combining charts via small multiples, in the form of trellis, lattice, grid, or panel charts, can help and is one of the best approaches available. However, these still only show a subset of these six critical dimensions or don't show them in context with one another.

3.2 Treemaps and other area-based visualizations

The ability of a treemap to represent multiple dimensions gives it more flexibility than, say, a bar chart, but mix effects continue to present problems. In fact treemap visualizations can be misleading if there are mix effects present in the data, a fact we believe has not been noted in the literature.

Recall that a treemap represents two dimensions of a data set, using rectangle size to represent one dimension and color to represent another [28]. Rectangles in a treemap typically represent a tree structure through their arrangements into groups and subgroups. A central intuition behind treemaps is that the values of individual elements "roll up" to show the values of overall areas. That is, by looking at the colors of a set of individual rectangles, a user can get a sense of the overall trend of the population.

Unfortunately, it's possible for this intuition to be strikingly wrong. A treemap in which the size dimension is not directly related to the color dimension is subject to all the problems of mix effects seen above. In a typical example of such a treemap, rectangles might represent companies, size might represent market capitalization, and color could represent change in profits. Due to mix effects, it's conceivable that all companies in a given market sector might show increased profits, even while profits have dropped for the sector as a whole–something that would likely surprise a typical user. In a large treemap with many subgroups, the chances that there's a Simpson's effect in one of the subgroups can be nontrivial.

Certain treemaps are safe from this problem, but still present difficulties when mix changes. In particular, some treemaps are explicitly designed to show changes in weights, using size to represent a dimension and color to show change in the same dimension. A well-known example is the SmartMoney Map of the Market [33], in which size portrays market capitalization of companies and color shows the percent change in stock price (generally equivalent to change in market capitalization). This type of treemap, where color is linked to the change in the size dimension, is sometimes called a *treemap for comparisons* [31]. It is immune to the purest form of Simpson's paradox, since if the overall value of a sector goes down, one of the items must have fallen in value as well.

Unfortunately, treemaps for comparisons still can be subtly affected by changes in mix. Consider a simple "market map" that portrays two companies with rectangles of equal area, one having gone up 25% and one down 25%. What has happened overall? In fact, the market has fallen by 6.25%: although the sizes of the companies are equal at the end of the time period, the company with the falling stock price had to have been bigger at the beginning of the time period. In other words, because treemaps for comparisons only show sizes based on the most recent data–that is, only half of the mix data–they subtly emphasize elements that have gone up in value, and thus have a systematic bias toward showing an increase in value.

These problems are not unique to treemaps. Any method where area is used to represent a dimension (mosaic plots, marimekko charts, and bubble charts) suffers from the same drawback. One potential way to fix this problem would be to base sizes on the values in the beginning period. This would have the obvious drawback, however, of not showing the most recent data. In other words, there's an essential tradeoff in treemaps between accuracy and timeliness when choosing the size coordinate.

3.3 Mix effects in maps

Unlike the case of treemaps, mix effects are a well-documented problem in the cartography literature [13, 30]. For example, imagine a state map of low birth weight rates by county, which showed highest rates concentrated in the east. How many viewers would realize that it's possible that when they drill down into state maps for each race, smoking/non-smoking, and gender combination for *every map* the highest rates are concentrated in the northwest [30]? Maps are especially susceptible to mix effects for two reasons: (1) their "size" dimension, geographic area, is often inversely correlated with population size and (2) confounding variables (race, age, etc.) are often spatially correlated with the variable of interest.

3.4 Related work on mix effects

In the statistics literature there are a variety of diagrams used to visualize mix effects, but we found few that meet our criterion of explicitly showing all six quantities described above. For example, consider the scatterplot from the original study on the Berkeley dataset, reprinted

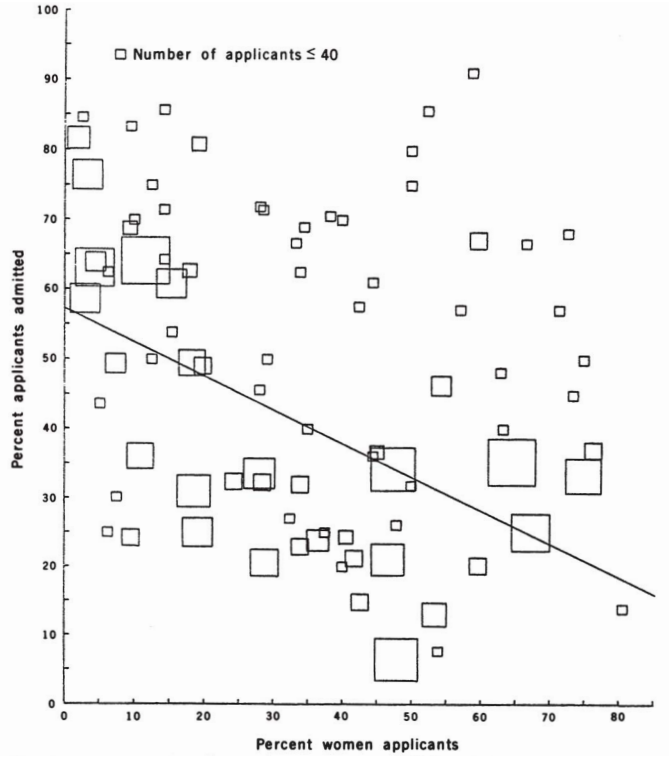

Fig. 2: A view of the Berkeley data from [8] (reprinted with permission). The diagram shows a negative relationship between department admission rates and proportion of female applicants for larger departments, which explains the "paradox." However, it provides no indication of overall or individual differences in admission rates.

in Figure 2 [8]. It clearly illustrates the crux of the famous Berkeley discrimination case, but nowhere does the diagram show separate admission rates for women and men. This is ideal for the purpose of this particular explanation, but is inadequate for a general tool of analysis.

Often diagrams of Simpson's paradox are created for pedagogical reasons. A good example is the beautiful interactive tool from Berkeley's Visualizing Urban Data Idealab [20]. This diagram, while clear and impressive, does not actually show all six quantities of interest. It also requires significant interaction to see the full story. As a result, it's not suited to a situation where an analyst needs to quickly understand a data set. The B-K diagrams [7, 18] similarly aim to clarify intuition, and are less suitable for analyzing complex data sets.

Two other examples that seem appealing at first glance are found in Agresti's textbook *Categorical Data Analysis* [5] and a graphic illustrating Simpson's effects on Wikipedia [26]. In these diagrams, for each subpopulation there is a rectangle or solid circle centered on a line, whose x position represents the value, and whose area represents the weight. Unfortunately, in a prototype system we quickly discovered this method has a serious drawback when used with larger numbers of segments. Large circles obscure small ones, and it's hard to see at a glance which subpopulations have changed most.

4 The Comet Chart

Given that comparisons of weighted sums and averages are a staple of scientific, business, and policy discussions, a natural design problem is to find ways to make mix effects transparent. Searching for a single visualization technique that is equally effective for all data sets and all users seems unlikely to succeed, so we narrowed our goal.

4.1 Defining the scope of the problem

We aimed to find a display technique that would help experts sort through problems related to mix effects. In particular, this project was motivated by the first author's experience as a senior analyst on a team of financial analysts at a Fortune 500 company. The author had done numerous painstaking analyses in which the conclusion was that a mix shift had driven top-line changes in critical metrics. A typical approach was to "slice and dice" the data, including using small multiples to see many slices at once, exploring interactively and iteratively. This was slow, required considering both size and value changes, became unrealistic for more than a small number of segments, and usually only caught big shifts while missing more subtle but meaningful patterns. The author wasn't alone: analysts in finance, product, engineering, sales, human resources, and marketing all worked on problems affected by changes in mix over time.

The finance analysts at this company were our targeted users. Such analysts don't necessarily have a degree in statistics, but all have a strong quantitative background and work with data daily. They investigate data in order to inform decisions made by leaders in sales, product, or finance. Previous analyses had revealed that mix issues were common and important. The analysts needed an easier, quicker, and more systematic way to diagnose situations where mix mattered.

Following the formalization of section 2, our goal is to visualize all six quantities that might matter to an analyst: starting and final values, starting and final weights, and the changes in value and weight. We assume that the analysts are in exploration mode, trying either to confirm that most segments are similar to the aggregate or to identify areas to analyze further.

To help analysts, a visualization of mix effects should provide fast answers to questions such as:

- Are the segments changing in similar ways to the aggregate? More simply, can we trust the aggregate?
- Are there extreme outliers, especially by weight, that can drag the aggregate and which we should isolate and analyze separately?
- Is there a relationship between value and changes in weight? Or between weight and changes in value?
- Are there subsets of segments that are changing in different ways, such that we should analyze them separately? Would identifying these differences lead to subset-specific decisions, treatments, or actions?

4.2 Comet Chart

To meet these needs, we created a *comet chart*. The starting point for the chart is the situation discussed in section 2: we assume there are two aggregate quantities of interest, P and Q, such that:

$$P = \sum_{i=1}^{n} a_i x_i$$

and

$$Q = \sum_{i=1}^{n} b_i y_i$$

The a_i and b_i represent weights and the x_i and y_i represent values for segments 1 to n segments, assuming the same set of segments in both scenarios. To represent all four of these dimensions at once, we use a variation of a scatterplot in which the x-axis represents weights and the y-axis represents values. For each i, we then draw a modified line segment from the point (a_i, x_i) to the point (b_i, y_i). The resulting "comet," flowing from tail to head, represents the change for the ith subpopulations between the two scenarios.

The "head" of the line segment is thicker than the tail, to create a sense of flow. Initially, we were inspired by Holten and Wijk's *A User Study on Visualizing Directed Edges in Graphs* [16]. Assuming the lessons from directed graphs might apply to disconnected directed lines, we started with the "needle" style taper they recommended. However, early user testing showed that the natural interpretation was inverted: people saw these disconnected shapes as "comets" or "fish"

rather than needles. In response, we kept the taper but reversed the direction to the untested "comet" type taper.

To make the comparison in aggregates clear, we also plot a line segment (in a markedly different color) connecting $(\frac{\sum_{i=1}^{n} a_i}{n}, \frac{P}{\sum_{i=1}^{n} a_i})$ to $(\frac{\sum_{i=1}^{n} b_i}{n}, \frac{Q}{\sum_{i=1}^{n} b_i})$. The x-coordinate represents the average segment weight. The y-coordinate represents aggregate value. In all, this distinguished comet, shown in black, represents the difference between the overall scenarios so that the aggregate can be compared to the individual segments.

The two examples in Fig. 1 illustrate how a comet chart works, using simulated data on unemployment rates and labor force size for six demographic groups. The left image shows an example in which the aggregate unemployment rate (the black comet) and the rates for each demographic segment of the population (gray comets) are moving in exactly the same way. In this simple scenario the size of the segments–i.e., the mix–hasn't changed, so the segment comets have no horizontal component and are colored gray.

The righthand image in Fig. 1 illustrates an example in which mix effects come into play. In this scenario too, the aggregate rate (black comet) has increased. However, the unemployment rate within each segment has remained constant! What has happened is that segments with high unemployment have grown (blue comets) while those with low unemployment have shrunk (orange comets). This is a Simpson's-like situation. The fact that the segments have changed only in weight is indicated by the pure horizontal direction of the segment comets.

That the aggregate is moving in a completely different direction from all segments may look strange–but that is exactly the point. It looks strange because Simpson's paradox feels strange, and the chart forces the user to confront the underlying issues. Color reinforces the positional change. In this example, the blue-orange color spectrum shows change in weight, reinforcing the horizontal shift and emphasizing the change in mix. Blue shows increasing weight relative to the total, orange decreasing, and grey neutral.

4.2.1 Color

Users can select from a range of color options to draw attention to different characteristics of the data: absolute change in weight, change in the "combined metric" defined by weight * value, increase or decrease in value, percent change in weight, etc. Color may be drawn either as a continuous spectrum showing a range of values, or a discrete orange/gray/blue scheme to show negative/zero/positive changes.

All color schemes use blue/orange, so that users with red-green colorblindness can perceive the difference. More subtly, an orange/blue scheme does not have a strong good/bad connotation. Since the comet chart may be used on many types of data, we did not want to mechanically assign a value judgment to positive and negative changes.

4.2.2 Scaling

Our target users frequently encounter log-normal or similar heavy-tailed distributions. As with many standard charts, allowing the option of a log scale for one or both axes proves helpful. It also allows us to compare either absolute or relative differences in value or weight.

Interestingly, there is an additional mathematical benefit of a log scale in the particular case of a comet chart. In the standard linear-linear scale we can't directly compare slopes. The same percent change in value or weight will be a larger distance on the chart if the original value or weight is larger. If we switch to a log scale on both axes, then the perceived slopes are given by the $log(percentchangeinvalue)/log(percentchangeinweight)$ and are therefore comparable. This is derived in the equation below.

$$\frac{log(size_1) - log(size_0)}{log(value_1) - log(value_0)} = \frac{log\frac{size_1}{size_0}}{log\frac{value_1}{value_0}}$$

There is a second, subtle advantage of a log scale. Level curves of the "combined metric", weight * value, correspond to simple diagonal lines $\log(weight) + \log(value) = c$ for any constant c. Because of this, for any given comet chart there's a straightforward relationship between the angle of a comet's tail and the change in the combined metric. While the exact relationship depends on the axis scales, an experienced user could use this as an extra cue during analysis.

4.2.3 Filtering

The tool enables filtering on certain criteria: increase or decrease in value, increase or decrease in weight, or increase or decrease in the combined weight * value metric. Comets that don't meet the desired criteria are shown nearly transparent.

By combining these criteria, one can show only those comets, for example, that have decreased in value *and* increased in weight *and* decreased in the combined metric. This may reveal groups of segments with similar characteristics.

4.3 Early design ideas

To arrive at the comet chart, we prototyped a series of ideas. Some included versions of the previous work discussed in section 3.4. For example, we worked with a design much like Agresti's with circles showing weight, position on a line for value, and lines drawn between the same circles on two scenarios to show change in value [5]. Unfortunately, with data sets larger than just a few items, this led to busy, unreadable displays in which weight changes were difficult to compare meaningfully. A second prototype used a connected scatterplot, showing more than two scenarios on a single screen. For more than a few segments, however, this ended up creating an overwhelming level of visual clutter and lost any sense of directionality.

A third straightforward idea is to generalize a bar chart, with one rectangle for each subpopulation. The height of each bar might reflect the aggregate value of that population, and (unlike a traditional bar chart) the width of each bar could show the size of the subpopulation. Unfortunately, this idea too is unwieldy in practice, due to the need to show two different subpopulation sizes and heights (one for each scenario). Sketches that showed both at once were uniformly judged to be unintelligible, especially in the common cases where changes in size or value were relatively small compared to the absolute numbers.

5 Examples: Tool in Action

5.1 Revealing mix shifts

How does complicated real data compare to our idealized situation? To demo the comet chart "in action" we chose a dataset that compares changes over time rather than differences between categories, as this best reflects the type of data relevant to our target analysts.

The next two examples seem quite similar to each other: both compare unemployment rates and the size of the labor force, segmented by education, for two different years [3]. However, the two comet plots reveal different economic factors. The first (Fig. 3) is much like our first idealized situation: all comets are flowing together. The second (Fig. 4) is like the second idealized situation: the aggregate's change is clearly different from the segments. In *both* we see that the labor force is shifting towards education categories with lower unemployment rates. This shift affects how the segment-level changes in unemployment rate relate to the aggregate changes.

In Fig. 3 we compare 2009 to 2013 and see that the unemployment rate has decreased for all levels of education as well as the aggregate. This confirms the naive assumption that the aggregate's decline is representative of the segments, which generally makes sense given the economic recovery since the recession in 2009.

But the chart also reveals an important second story about the changing demographics of the labor force during this time. The slope and color of the comets shows that education categories with higher unemployment rates are *decreasing* as a percent of the labor force (comets slanting left and colored orange) while those education categories with lower unemployment rates have an *increasing* labor force (comets slanting right and colored blue). The country-wide unemployment rate drop is driven both by decreases in unemployment within each segment *and* the fact that the labor force is more highly educated than four years earlier. Without the comet chart, the latter factor would be easy to overlook since the changes in subpopulation unemployment rates are directionally the same as the aggregate.

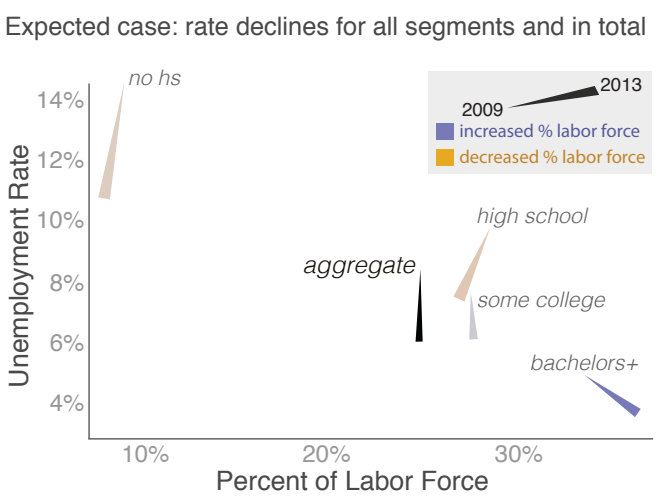

Fig. 3: The unemployment rate decreased both in aggregate and for all education levels. *X*-axis: percentage of the labor force ("weight"). *Y*-axis: unemployment rate ("value"). Additionally, low education and high unemployment segments shrank (orange color and horizontal shift to the left show decreasing weight) while the high education and low unemployment segments increased (blue color and horizontal shift to the right show increasing weight) as a percent of the labor force. Color is continuous, defined by change in weight.

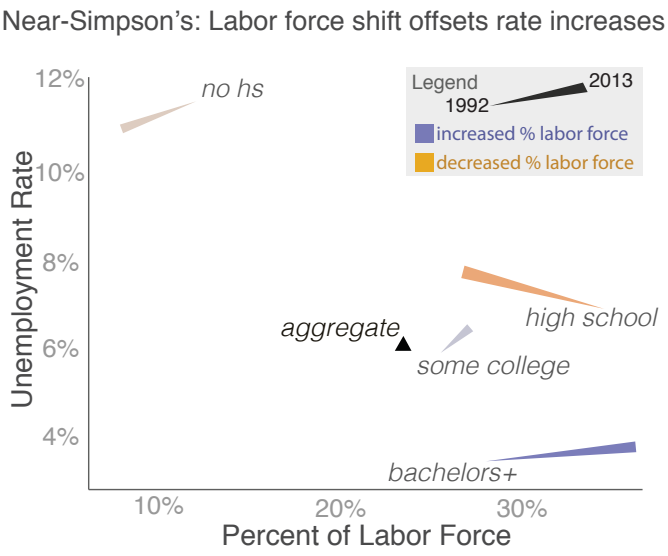

Fig. 4: A near-Simpson's effect. Aggregate value is nearly constant, while all segments change in both relative size of the labor force and unemployment rate. Higher education segments increased as a percent of the labor force while less educated segments shrank. The weight shifts offset the increases in unemployment for the three largest categories. Color is continuous: defined by change in weight.

The next example in Fig. 4 compares 1992 to 2013, both shortly after a recession. Like the first example, we see a relationship between unemployment and changes in the size of the labor force. However, the changes in unemployment rate are now much smaller compared to the changes in relative size of the labor force. In fact, the three largest segments show *increasing* unemployment rates while the aggregate shows a very slight decline. Just looking at the value changes, it might be tempting to assume that the aggregate decline is driven primarily by the decline in unemployment for those without a high school degree. But, in the comet chart, it's clear that this segment is much too small to offset the increases of the larger segments. The fact that the comets are mostly horizontal rather than vertical shows changing mix. Essentially there is a trade-off between the increasing size of the labor force for the categories with lower unemployment rates driving unemployment down, offset by the increases in unemployment rate within each large category.

In other words, the horizontal "motion" indicates that understanding mix is critical to analyzing this data. However, unlike our idealized dataset in which the aggregate changed in just value and the segments changed in just weight, in this example the aggregate metric shows almost no change while all segments change in value *and* weight. While we might want to dig deeper to learn exactly what is going on, it's certainly clear that the aggregate is not representative of the segments and that the changing demographics of the labor force are important to understanding the relationship between wages and education.

5.2 Disaggregation is valuable

Even without a Simpson's-like effect, visualizing disaggregated data with the comet chart is still useful to highlight outliers, show when segments are behaving similarly or differently, suggest relationships between the same 6 key quantities, or reinforce/refute the base hypothesis that segments are changing similarly to the aggregate. This is especially helpful when there are many (more than 20) segments, which was typical of the data analyzed by our target users. To illustrate this, let's look at the year over year change in unemployment rates and labor force size for counties in four US states[3] [1].

First, like a scatterplot, the comet chart reveals outliers in terms of unemployment rate, as in Imperial County in California, or in the size of the labor force, as in Los Angeles in California (Fig. 5). Moreover, the comets also show outliers in terms of the *change* in the unemployment rate, as in Hale, Swisher, and Floyd counties in Texas (Fig. 8).

Next, for California (Fig. 5) and Ohio (Fig. 6) the chart reinforces our naive assumption that county-level changes are generally directionally similar to the state-level changes. In contrast, in Michigan (Fig. 7) and Texas (Fig. 8) the state unemployment rate is almost unchanged, belying potentially important county-level information.

For example, the chart for Michigan (Fig. 7) suggests a possible relationship between the size of the county and the increase or decrease in unemployment rates. A few large counties decreased their unemployment rate, as shown in orange on the right, while the unemployment rate increased in almost all of the small counties as shown in blue on the middle and left. Using a log scale on the x-axis as shown, and selecting the appropriate filter and color scheme can make this more discoverable. Obviously this type of relationship between county size and changing employment rates might have important policy or economic implications for the state and is likely worthy of further study.

5.3 One last look at the original median wage example

Returning to our original example from the *New York Times*, using the comet chart as shown in Fig. 9 we can now see the mix effects at play in the median wage data.

6 Deploying the new technique

To deploy this new technique, we built it as part of a more comprehensive tool (Fig. 10) in which analysts input data, select color, scale, or filter, hover on or select comets to get more details, and see the same data in a supplementary table. These elements make the visualization technique accessible and enable analysts to create and explore comet-type charts with their own data.

6.1 By user request: a sortable table

The first version of the tool had only the comet chart. Test users asked for a complementary table showing segment names and exact values. Adapting the "sortable table with bars" from the popular D3 library example [9], we added a sortable table with six mini-horizontal bar charts (Fig. 10). The bar charts proved useful as a supplement to the comet chart and as a standalone feature.

[3] comparing Sept 2012 and Sept 2013

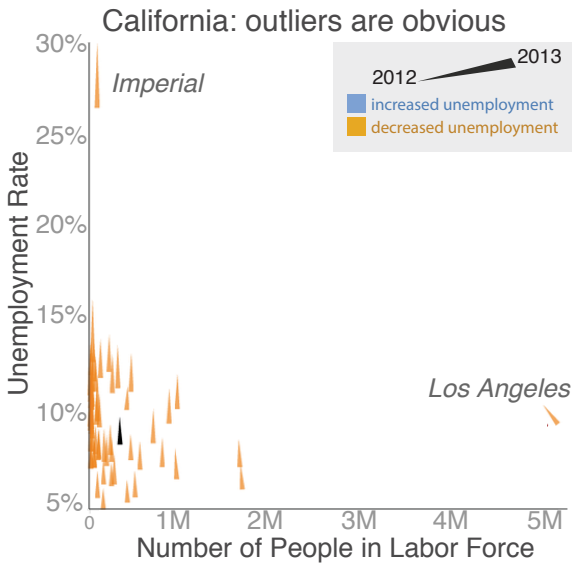

Fig. 5: California: Unemployment rate decreases overall and in most counties. However, the chart reveals outliers: Imperial's high unemployment rate and Los Angeles' large population of laborers. Color is discrete, defined by change in value (unemployment rate).

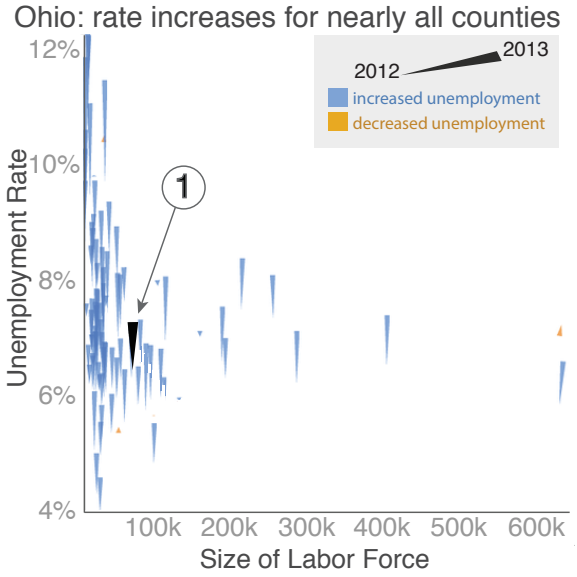

Fig. 6: Ohio: Unemployment rate increases overall (1) and in most counties. The power-law type distribution of county size is typical: toggling between a linear and log-scale on the x-axis helps users explore this type of distribution. Color is discrete, defined by change in value.

6.2 Form of the data

Despite the wide range of applications, the data required for the comet chart tool can be expressed in a simple comma-separated text format that our users found intuitive. Users enter data either by copying/pasting from a spreadsheet or submitting a special query identifier to import data from a commonly-used internal database. The latter option proved to be one of most valued features of the tool.

7 User Feedback

7.1 Process for gathering feedback

While the comet chart is a general technique, the tool was specifically designed to help financial analysts at a Fortune 500 company quickly visualize disaggregated data and identify if mix mattered for key business metrics. We gathered qualitative feedback by sitting with experts as they used the tool or, if out of town, by asking them to demo and send feedback. Interviews included observation, answering questions, asking questions about what they were doing or noticing, discussing interpretations and discoveries, and asking for desired features. There were two rounds of feedback: an initial round to inform chart and tool development and a second round to validate if the tool worked in practice. In total, we received feedback from 28 individuals with relatively little overlap between rounds.

The first round of interviews occurred during development of the tool and helped inform its design. This included conversations with 15 people, including 9 likely target users and 6 others who would have a valuable perspective as statisticians or senior analysts, familiarity identifying mix effects from earlier work, or as a fellow tool-creator. 3 target users provided their own data, while relevant datasets were supplied to the other 6. 5 sessions were one on one, while 1 was with a team of four who worked closely together. Most conversations were in-person over the course of one week in the company's home office, while we followed up by email or video conference with 4 others who worked in other offices including abroad.

After the initial launch, we introduced the tool with a wider group of potential users via introductory presentations and workshops. We then reached out to known users for a second round of interviews and with a survey. The interviews included qualitative one-on-one in-person sessions with 3 analysts, and session with a group of 4 from another department. All interviewees provided their own data. Additionally, 11 people replied to the survey, 9 who had input their own data and 2 who had used the provided business-specific demo data. 10 were active analysts (9 in finance, 1 in marketing) while 1 was a software engineer. These 11 represented 9 different work-groups in 5 offices, and on 3 continents.

While our evaluation elicited useful practical feedback, it's important to note some of the drawbacks compared to a formal study. The fact that those opting to use the tool are self-selected and many are known personally to one of the tool-creators runs the risk of positively biasing our results. At the same time, these colleagues are not shy to offer critique. They're also busy and unlikely to waste time on something if it's not helping them. Following Sedlmair et al [27], our view was the benefits of working with real target users in their own environment counterbalanced the potential pitfalls and subjectivity.[4]

7.2 First round of feedback: informing the design process

During the first round, our goal was to learn how the comet chart worked with real users, real data, and real analytic questions to decide if we should pursue this further or try something else. The results were promising. Analysts engaged with the tool actively; one said *"this is crazy fun!"* They generally confirmed they saw things they expected to see and also discovered new things.

At the same time, we heard helpful critiques as well which are included in section 7.5. In particular, this feedback helped us to clarify the scope of the tool.

7.3 Post-deployment survey

In our post-deployment survey, we sought to find out if the analysts were aware of and needed to understand mix, how they currently addressed this need and if their current methods were satisfactory, and if the comet chart tool met this need well. Our respondents' answers showed a need to understand mix and that current tools are not effectively meeting this need. 10 out of 11 respondents said that mix issues come up "sometimes, often, or very often" in their work. Despite being aware of mix issues and encountering them regularly, 5 said that their current tools reveal mix issues "not at all well or slightly well" while 5 others said "somewhat well" and only one said "very well." In contrast, 9 respondents said the new comet tool worked "very well" or

[4] "personal involvement of the researcher is central and desirable, rather than being a dismaying incursion of subjectivity that is a threat to validity [27]."

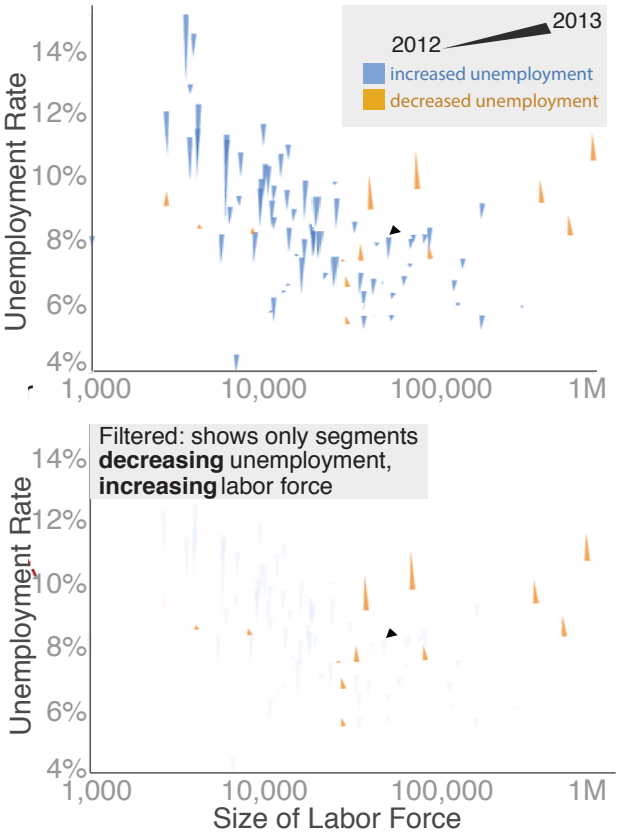

Fig. 7: Michigan: Stability in aggregate unemployment rate belies county-level movement. Rate decreased in three largest counties (shown in orange, right side of chart); it increased in nearly all small and mid-size counties (blue). This relation between county size and change in unemployment rate is likely of interest. Color is discrete, defined by change in value.

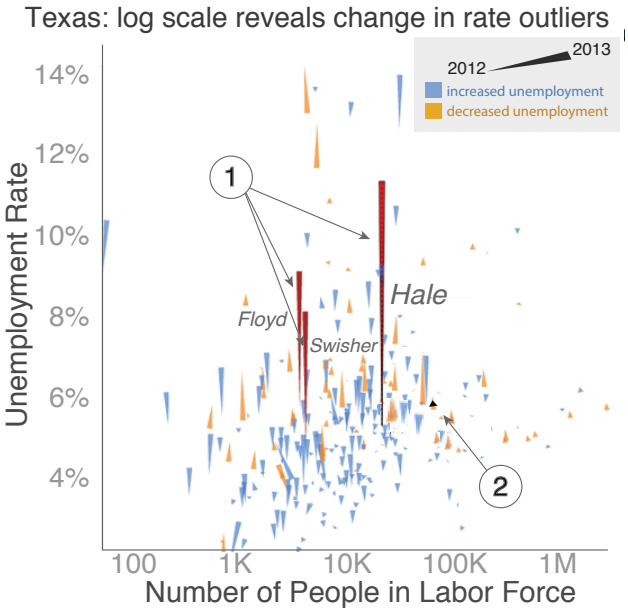

Fig. 8: Texas: An *x*-axis log scale focuses attention on numerous small counties. Three counties show a dramatic increase in unemployment rate (1). These three have been selected, so comets are shown in red. In fact, one county had experienced a manufacturing plant closure; the other two are geographically adjacent. Aggregate is stable (2). Color is discrete, defined by change in value.

"extremely well." One user summarized it saying that *"I don't have another quick way to visualize mix effects that could be taking place."*

7.4 The comet chart and tool are useful

One user described the tool as an *"intuitive way of visualising all the moving parts within a population"* while another explained that it *"was simple to use and incorporate one's own data."*

A third analyst wrote: *"Comets goes a long way to solving a problem that often goes unnoticed: aggregate metrics rarely tell the whole story. Countless times, presentations treat the mean as if it represents the whole. In an effort to simplify, we reduce data about hundreds of pieces to a single number."* His actions supported his words, as we found him using the tool late one evening for an urgent analysis.

In at least one case, the tool helped change a decision: an analyst explained *"I could see a striking change in mix, which invalidated high-level goal setting .... I needed to go back and set goals based on the current composition."*

7.5 Pros, cons, and areas for improvement

To summarize, our initial deployment showed that the comet chart provided value to many of our users. At the same time, feedback revealed that the comet chart handled certain cases better than others, and that there are areas for improvement.

7.5.1 Data

Comet charts work best for certain types of data. The strong sense of motion conveyed by comets worked well for time-based comparisons, but consistently confused users when there was no inherent order to the scenarios. For some users this was an inconvenience, while a deal-breaker for others.

Other testers noted that datasets with different numbers of segments needed different design solutions; for them, the comet chart worked best for 6-100, was weaker for over 100, and failed over 500. Too few segments resulted in a poor data to ink ratio and simpler tools like a table would suffice. Too many segments resulted in occlusion or challenges due to extreme outliers. In practice, we found that there was a great need for visualizing data in this middle "goldilocks" space of 6-100 segments. Analysts found that this technique also worked for long-tailed distributed data, which they faced often and which seem especially prone to mix effects since different factors may be causing meaningful changes to size and value of head, torso, and tail segments.

7.5.2 Design

Some users questioned the "comet" device for showing direction of line segments. Early on, we switched from an "arrow" to a "comet" interpretation based on feedback from users that "comets" matched their natural intuition. Nonetheless, we found first-time users often checking the legend, some finding the direction more natural than others. The issue generally resolved as users learned the convention.

More generally, learning what to look for and how to interpret the chart quickly does take some time and experience, as it would for learning to interpret any new type of chart. One analyst confirmed, however, that it did make sense: *"In a clear case, where the comets are all over the place, the message is brought home very elegantly in the chart."*

A final issue is that long comets draw more attention than short ones. When these changes are the key issue, or when comparing changes across segments with different weights or values, this is an advantage. However, some users pointed out that segments with large changes may attract undue attention while smaller changes might be harder to see. Similarly, it can be hard to see the aggregate comet when its value and weight haven't changed significantly relative to the chart scale, because then the corresponding black comet is small. One user suggested adding a prominent text description of the aggregate

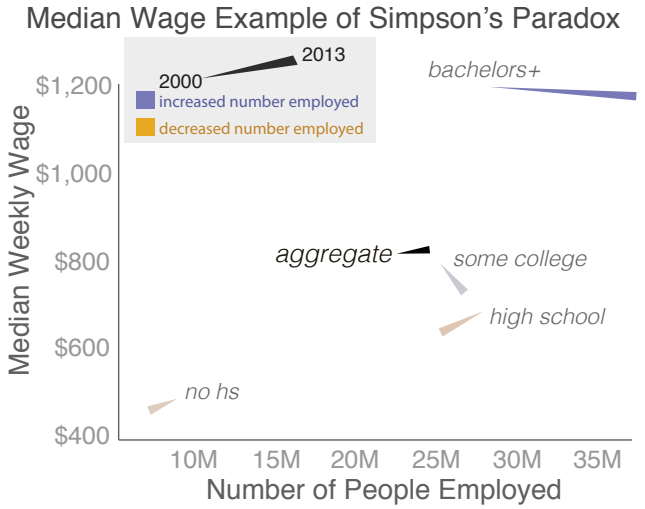

Fig. 9: The original median wage example: the strong horizontal motion, especially in "bachelors+", indicates important mix shifts. The aggregate's slight upward tilt, showing an increase in median wage, differs from the downward trajectories of all subgroups. The strong horizontal movement emphasizes the importance of the weight changes. Highly paid categories are growing, while low-wage categories shrink. In aggregate, the shift in the labor force towards higher education/higher wage segments offsets the declining wages of each segment. Color is continuous, defined by change in weight.

change.

7.5.3 Exploration and interactivity

A desire for additional interactive and exploratory capabilities was a consistent theme. As one analyst put it, changes to the visualization itself would have *"not as much benefit as flexible axis, multiple charts, more filters, etc."* A few users chose not to restrict themselves to data that fit the "value-weight" construct, and used the tool to more generally explore change over time in two different metrics. Many of the requested features were ones which would aid exploration, like more complementary charts. Or, the ability to use comet charts to more easily explore more complex datasets: more than 2 points in time, multiple dimensions, or a hierarchy of dimensions to use as segments.

7.5.4 Sharing and collaboration

In early testing, several users confirmed that our likely audience was analysts, skeptical that it would work for presenting results to business leaders. However, once we had deployed the tool and analysts started making discoveries, the top feature request was easy ways to share what they found with non-analyst partners: emailing a URL, converting to a "presentation-ready" format, or embedding in a dashboard.

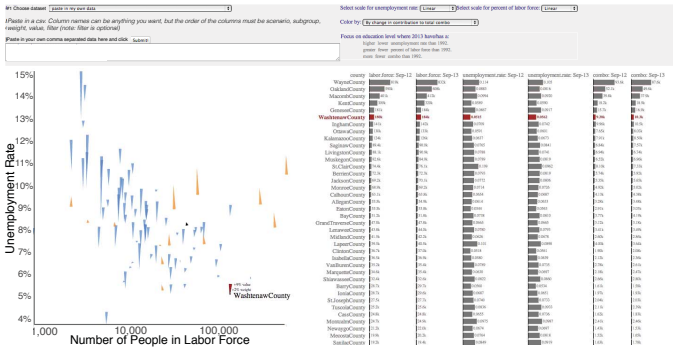

Fig. 10: Comet chart embedded in the tool, supplemented by a sortable bar chart view (right) and filtering/scale/highlighting options and data input (top). Bar chart and comet charts are linked: one county is shown in red on both the comet chart and bar chart in response to a click.

8 Conclusion and Future Research

Mix effects are ubiquitous and confusing. They can surprise and mislead lay users of visualizations, and even expert analysts can benefit from better tools to understand their implications. As we have seen, common business charts and maps make it difficult to show the effects of simultaneously varying subpopulations' size and metrics. Furthermore, we have shown how more sophisticated tools that can show multiple dimensions, such as treemaps, mosaic plots, and bubble charts, are nonetheless susceptible to perceptual biases.

Just as there is no statistical silver bullet to account for mix effects, there is likely no single magic visualization that makes them obvious. Instead, it is the responsibility of analysts to look for these effects and for information designers to make sure that any visualization of aggregate data makes it clear how overall data relates to subpopulation metrics. One purpose of this paper is simply to heighten awareness: we believe that many visualization designers and users underestimate the challenges of presenting aggregate statistics. Despite the many treemaps published on the web, for example, we know of none that attempts to correct for the upward bias caused by changing sizes. At a more prosaic level, it's common to see visual displays (as in the *New York Times* article) that omit any reference to mix.

We believe that there is room for a portfolio of new techniques that can help designers and analysts alike. The "comet" visualization presented in this paper is an example: a straightforward modification of a scatterplot that displays the interaction between subpopulation sizes and values in a balanced view. It's designed for analysts who want to understand complex data. Based on reactions from analysts at our company, we believe the technique has promise. Future work could make it more valuable by enabling more exploration and more complex datasets.

Of course, the comet chart is tuned for a specific use case and audience: an analyst looking at aggregates of many items, comparing two different time periods. Changing any of these conditions could lead to an opportunity for a new technique. Could there be an easy way to compare multiple time periods and scenarios? What if there are new segments or dropped segments? Are there techniques that might work for aggregates of just a few items or perhaps of hundreds or thousands? How can we compare categorical scenarios rather than temporal?

In particular, it would be extremely useful to find ways to present mix effects to non-experts. Expert analysts usually need to explain their insights to business leaders, who may not have a statistical background; journalists need to write stories that can be widely read. Because the comet chart requires a learning curve, it is not optimal for a casual user; so this remains an important area for future investigation. Given that basic mix-effect situations involve as few as eight numbers (two values and two weights, across two scenarios), finding a clear way to present them is a crisp, simply-stated research problem.

In summary, current charting tools are not meeting the needs of users, experts and amateurs alike. At the same time, visualization designers often do not take into account the potential surprises that mix effects can cause. We've described one method, the comet chart, for handling these issues, and we believe that finding ways to address mix effects in visualizations is an important and promising area for future research.

Acknowledgments

The authors wish to thank Danielle Romain and Jon Orwant for making this research possible. Fernanda Viégas provided critical perspective and advice. We thank Alison Cichowlas and Colin McMillen for useful feedback and technical help. We also thank Eric Tassone and Bill Heavlin for their statistical expertise and our many testers for trying the tool and giving feedback. We thank the anonymous reviewers for their helpful critiques.

REFERENCES

[1] Bureau of labor statistics: Local area unemployment statistics. labor force data by county, not seasonally adjusted. http://www.bls.gov/lau/laucntycur14.txt, November 2013.

[2] Bureau of labor statistics: Current population survey labor force statistics - table 5. quartiles and selected deciles of usual weekly earnings of full-time wage and salary workers by selected characteristics, quarterly averages, not seasonally adjusted. http://www.bls.gov/webapps/legacy/cpswktab5.htm, January 2014.

[3] Bureau of labor statistics: Current population survey labor force statistics - table a-4. employment status of the civilian population 25 years and over by educational attainment. http://www.bls.gov/webapps/legacy/cpsatab4.htm, January 2014.

[4] Bureau of labor statistics: Inflation adjustment calculator. http://www.bls.gov/data/inflation_calculator.htm, January 2014.

[5] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.

[6] O. A. Arah. The role of causal reasoning in understanding simpsons paradox, lord's paradox, and the suppression effect: covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology*, (1):5, 2008.

[7] S. G. Baker and B. S. Kramer. Good for women, good for men, bad for people: Simpson's paradox and the importance of sex-specific analysis in observational studies. *Journal of women's health & gender-based medicine*, 10(9):867–872, 2001.

[8] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975.

[9] M. Bostock. Sortable table with bars. http://bl.ocks.org/mbostock/3719724, 2012.

[10] C. Charig, D. Webb, S. Payne, and J. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British medical journal (Clinical research ed.)*, 292(6524):879, 1986.

[11] K. A. Clarke. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352, 2005.

[12] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, Inc., 3rd edition, 2003.

[13] A. Gelman, P. N. Price, et al. All maps of parameter estimates are misleading. *Statistics in Medicine*, 18(23):3221–3234, 1999.

[14] D. L. Guber. Getting what you pay for: The debate over equity in public school expenditures. *Journal of Statistics Education*, 7(2), 1999.

[15] S. J. Hahné, A. Charlett, B. Purcell, S. Samuelsson, I. Camaroni, I. Ehrhard, S. Heuberger, M. Santamaria, and J. M. Stuart. Effectiveness of antibiotics given before admission in reducing mortality from meningococcal disease: systematic review. *BMJ: British Medical Journal*, 332(7553):1299, 2006.

[16] D. Holten and J. J. van Wijk. A user study on visualizing directed edges in graphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2299–2308. ACM, 2009.

[17] W. L. James, R. E. Cossman, J. S. Cossman, C. Campbell, and T. Blanchard. International journal of health geographics. *International journal of health geographics*, 3:7, 2004.

[18] J. Jeon, H. Chung, and J. Bae. Chances of simpson's paradox. *Journal of the Korean Statistical Society*, 16(2):117–127, 1987.

[19] S. A. Julious and M. A. Mullee. Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481, 1994.

[20] L. Lehe and V. Powell. Simpson's paradox: Girls gone average. averages gone wild. http://vudlab.com/simpsons/.

[21] F. Norris. Median pay in U.S. is stagnant, but low-paid workers lose. *New York Times*, Apr. 2013.

[22] L. W. Pickle and A. A. White. Effects of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine*, 14(5-7):615–627, 1995.

[23] M. L. Radelet and G. L. Pierce. Choosing those who will die: Race and the death penalty in Florida. *Fla. L. Rev.*, 43:1, 1991.

[24] R. Ralf, B. Annette, P. W. van, and G. J. Mintjes-de. Simpson's paradox: An example from hospital epidemiology. *Epidemiology*, 11(1):81–83, Jan. 2000.

[25] K. Ross. *A mathematician at the ballpark: Odds and probabilities for baseball fans*. Penguin, 2007.

[26] Schutz. Simpson's Paradox.svg — Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/File:Simpson%27s_paradox.svg, Aug 2007. [Online; accessed 12-March-2014].

[27] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2431–2440, 2012.

[28] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, Jan. 1992.

[29] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241, 1951.

[30] E. C. Tassone, M. L. Miranda, and A. E. Gelfand. Disaggregated spatial modelling for areal unit categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(1):175–190, 2010.

[31] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1121–1128, 2007.

[32] R. L. Wardrop. Simpson's paradox and the hot hand in basketball. *The American Statistician*, 49(1):24–28, 1995.

[33] M. Wattenberg. Visualizing the stock market. In *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '99, pages 188–189, New York, NY, USA, 1999. ACM.

[34] G. U. Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903.