CS 171: Process Book
Ballroom Events Explorer

Ross Rheingans-Yoo

I. Overview / Motivation

I've been dancing with the Harvard Ballroom Dance Team for a year and a half, and this year, I've served as the team's secretary. In that time, I've traveled to more than a dozen collegiate competitions with HBDT, each of which includes several hundred competitors, hundreds of rounds of competition, and thousands of judging marks (which are, conveniently, posted online).

This year, I've been working with the team's executive board to better understand our team's performance at competitions (to inform coaching/training decisions, mostly), which is no small task, as the Harvard team is sometimes in excess of a hundred dancers! To date, my methods have included taking notes manually during/after a competition, compiling the data in Microsoft Excel or similar, and using rudimentary graphing tools to explore trends.

I plan to apply my newfound skills with D3 to create a few dynamic visualizations that draw from competition results scraped from online sources, aiming to answer a few questions that the HBDT executive board has been asking recently.

II. Related Work

While online reporting of competition results is relatively new in collegiate ballroom dance, ballroom dancer and web developer Cloud Cray has done some work writing a web-scraping toolkit for O2CM[1], one of the most popular registration/results websites. This work[2] will save me a lot of difficulty in compiling /parsing results, and allow me to focus on visualizing and presenting the actual data.

Mr. Cray has been running some statistical analysis and published his findings as an infographic on his blog[3], but has mainly focused on large-scale trends across the US. By contrast, I'd be interested in exploring only a handful of competitions (and a handful of teams) at a time, which, nonetheless, will still represent a significant corpus of data.

III. Questions

After discussion with a few dancers and other members of the HBDT executive board, as well as Mr. Cray, I've identified a few interesting questions:

- Are the recent changes in coaching working? (Are we getting better between years?)
- What particular styles/dances/events are we best/worst in?
- What correlation exists between competitive performance and long-term commitment to HBDT? (Is early competitive success a good predictor of which dancers will stay on the team?)

The last, of course, requires cross-comparison of competitive results with outside information (HBDT membership rolls), and is probably outside the scope of this project, but the other two could be addressed by a graphical interface for tracking the performance of selected dancing couples across multiple rounds of competition, and comparing results across events/competitions/years.

[1] http://www.o2cm.com/

[2] https://github.com/CloudCray/DanceMarkScraper

[3] http://i.imgur.com/gJBUpJJ.jpg

Additionally, I expect that the visualization might see use by dancers, captains, or coaches wishing to understand competitors' results (their own, or others') in larger context of a round, event, or competition

IV. Data

A. Acquisition / Processing

I scraped competition results data scraped from O2CM, using the previously-mentioned DanceMarkScraper. This gives me access to almost all competitions attended by HBDT in the past several years, as O2CM is by far the most popular competition results reporting site in the US. I considered writing additional scraping tools to scrape similar data from other results reporting sites, but weighing the effort required to process their (often fragmentary) data into the forms available fromt O2CM scrapes against the number of additional competitions that it would make available, I determined that it's not worth the effort, and that I intend to focus solely on data scraped from O2CM, which covers all but a handful of competitions Harvard has attended in the past several years.

The final scrape was performed on April 30, 2014, and collected a corpus including data from 297 competitions between 2007 and 2014, totaling $\tilde{4}$.2 million judging marks for $\tilde{3}$00,000 competitor couples.

From there, I used my own Python scripts to aggregate the raw data into a hierarchical format (events/competitors/rounds/results) and write it out as JSON for use in the web front-end, some further data processing takes place ad-hoc within d3.js scripts, client-side. Wrangled and written out as CSV and JSON, the dataset is $\tilde{6}$20MB in total (both flat and aggregated).

B. Format

I'm using some jargon-ish terminology in describing the structure of competition results data, so it might be useful to describe the rough format of ballroom competition, and how results data are structured.

Within one *competition*, several dozen *events* are made up of several (between one and six) *rounds*. (Confusingly, both events and rounds are refered to as *heats*.) Each event may be contested by up to several dozen *couples* (or competitors). In each round up until the final round, several judges will make a judgement about each couple — voting either to recall that couple to the next round or not. If the number of votes to recall exceeds some threshold (different round-to-round and between events), the couple is recalled to the next round; otherwise, they are eliminated and will stop dancing. In the final round, a more complicated system is used to sort all finalist couples in order.

Thus, couples' results in an individual event are (roughly) based first on the number of rounds reached, and second on the number of recall votes (or *marks*) received in their last round danced. The system, of course, is slightly more detailed than this in practice, but this explanation at least explains the structure of the data I'm visualizing.

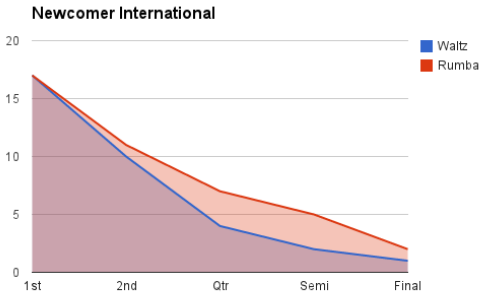
V. Exploratory Analysis

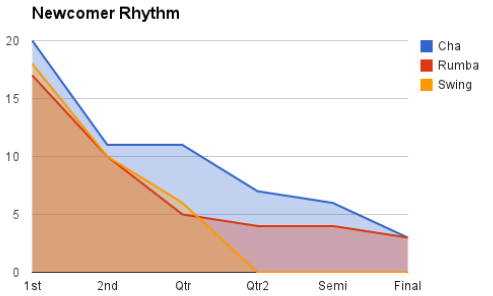
I've performed relatively little exploratory data analysis thus far, as my efforts in building the first prototype have mostly focused on the ability to display results data from a single event in an interactive format. In the next week, I expect to use this single-event view to perform exploratory analysis in comparing different events (across levels, styles, competitions, or years) in order to discover interesting trends to visualize.

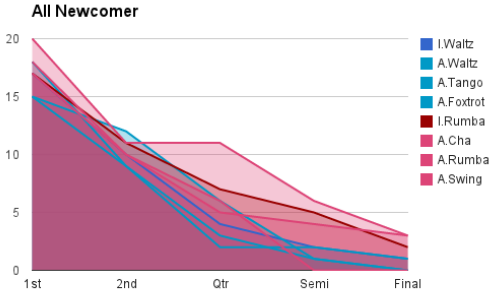
VI. Design Evolution

A. Pre-design

The initial idea for this project began with some rudimentary graphs I made over winter break in Microsoft Excel based on competition data collected manually this past fall:







Here, I was graphing the total number of Harvard competitors remaining in the competition after each successive round. (In the second graph, for example, note that Harvard dancers did much better in Rumba than in Swing in the Rhythm style, but only in later rounds, whereas they did better in Cha Cha than Rumba early on, but the two ended near the same in final results.)

Despite (or perhaps, due to) its conceptual simplicity, this can be a powerful visualization for comparing the aggregate results of small groups against each other (either the same group in different events, two groups in the same event, or the same team across different competitions or years), and I've played around with graphing some of these comparisons. However, the labor required to manually look up and input results quickly became prohibitive.

B. Initial Design

When I took up this project for CS 171, I built on some of these preliminary ideas using design principles learned in class. In particular, I needed to be aware of the principle of maximizing data-to-ink ratio, the effective use of color, and the concept of detail-on-demand. While I still liked the overlaid-area-graph concept, I wanted to add a stacked-graph option (for comparing one or more teams against the field at large in a single event).

Of course, the possibility for interactivity represents a significant addition to static graphs, and I wanted to make effective use of this capability in d3. However, I didn't have a concrete idea of how I wanted to incorporate interactive elements until Design Studio; I've described the design evolution of the interactivity dimension below.

C. Design Studio

I completed Design Studio #3 in class on 10 April 2014, with Jack Davison (the rest of his group was otherwise occupied), and peer-reviewed his group's project on visualizing trends in Q guide data. We shared both design sketches and preliminary prototypes for critique and discussion. Several points that we discussed shed light on important design considerations in my project, so I've listed the major ones below:

1) Audience: In discussion, Jack and I realized a fact that might be useful for both of our projects: our prospective userbases were made up of two relatively-distinct types:

- top-down users, who have relatively little domain-specific knowledge, but are browsing for interesting trends.
- bottom-up users, who have domain-specific knowledge and are looking to

draw connections between known facts and additional interesting data features.

In my case, specifically, the former corresponds to users who don't know (or don't care) about individual competitors, but are interested in aggregate statistical results, and the latter corresponds to users who know some handful of competitors by name, and want to explore their results in the context of the field at large.

As these two have different interests and needs, it's important to consciously consider and balance the needs of both groups. Specifically, taking this into consideration led me to place a greater focus on interactivity as a method of revealing data to bottom-up users who might be interested in it.

2) Data-to-Ink Ratio: One of the major critiques Jack had of my design sketches was that their data-to-ink ratio was strikingly poor (or perhaps more properly, their data-per-square-inch ratio). He suggested subdividing the stacked area charts (of one or two groups) into smaller regions making use of color scales to increase data density.

A few premilinary computer sketches later revealed that this was a brilliant suggestion, and that a segmented bar-chart design is a fantastic way to display data where an individual competitor's result is made up of temporal data (what round reached) as well as an aggregate of preceding rounds. I expect to use this as a default view, and to use the previously-described area/stream graph as a secondary, toggle-able option.

3) Tabular Detail View: Ironically enough, as we discussed, the high-density, detail-focused view sparks a desire in users for yet more detail data. If a particular round appears unusual or interesting, a user will want the ability to explore it in more depth. Since we have the data already available, it makes sense to add a sidebar-based detail view that allows for comparison of competitors within a round. (Preliminary sketches seem to indicate that this information could be easily conveyed in a tabular format, not unlike Homework 1.) I will be exploring this possibility further in later prototypes.

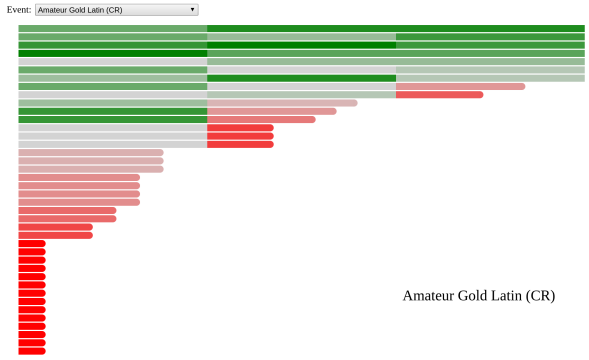
4) Overall area vs. Percentage area: Jack suggested that, in views comparing the number of competitors remaining from a selected group to the total, both absolute-scale and relative-scale measurements might be informative. As such, I decided to add a toggle-option to switch between absolute-scale (number of in-group competitors remaining, stacked with total remaining competitors) and relative-scale (% of in-group competitors, compared to total competitors, per round) in area-graph views of competitors remaining over time.

D. Final Design

VII. Implementation

A. "Competitors" view, single event

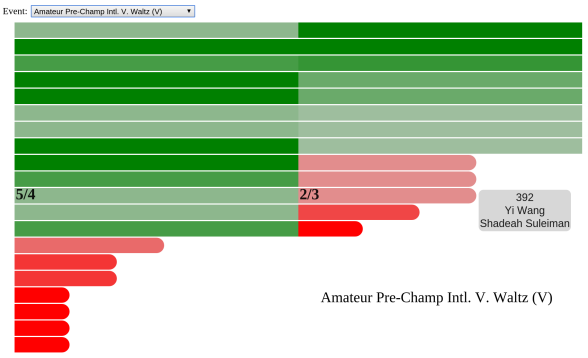
Based on feedback in design studio, I decided to use a relatively information-dense view that displayed all competitors in a particular and their individual round-by-round. A single 'pane' represents a single event in terms of all of its competitors, their round-by-round results, and overall placement.



Here, each bar represents a single couple competing in the event (here, Amateur Gold Latin Cha/Rumba), the length represents their success (large steps represent the number of rounds recalled; within a round, longer bar means closer to being recalled to the next round), color represents score within a particular round (on a divergent scale, with green/grey/red representing marks

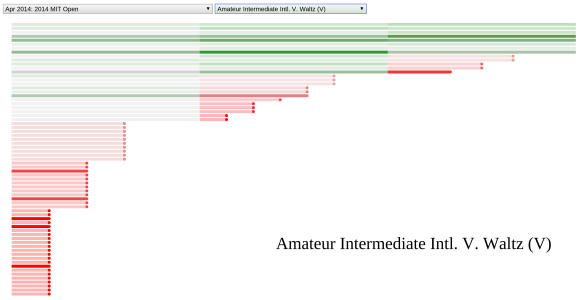
above/at/below the cutoff for recall), and vertical position simply sorted by final result.

On mouseover, more detail is revealed:



- tooltip: couple number, lead and follow's names
- round-by-round numeric score, compared to recall cutoff

Clicking on a bar highlights that bar, and the total set of selections is retained over switching the event or competition shown — this feature is useful for tracking a particular set of dancers across multiple events or competitions. (For dancers who dance with multiple partners, they are tracked across competitions by full name.) Pre-set selections of dancers' names can be loaded (currently, the only one available is a list compile from the Harvard Ballroom Dance Team's roster), and work in the same way as manual highlights.



For comparing different events, options at the top of the page allow the user to place several such panes side-by-side. A pair of dropdown menus (at top of each pane) allows, for each, the selection first, of any competition in the data set, and second, any event within that competition. In this way, the user can compare two or more events directly. Competitors highlighted in one pane are likewise highlighted in the other.

As picking highlighted dancers out of the whole list becomes difficult as the number of panes increases, a toggleable option allows the user to modify the sort function to place highlighted couples at the top.

VIII. Evaluation

As the data collected spans many years and dozens of separate competitions, I have not yet fully explored it. However, a few preliminary experiments have demonstrated its usefulness in answering questions that the HBDT Executive Board is interested in, regarding the team's competitive success over the years. Below, I've described one example of such a question where this visualization has been able to provide a useful answer that would not be apparent otherwise.

A. Rookie Coaching: '12-'13 vs. '13-'14

At the end of last year, the HBDT Executive Board was forced to make budget cuts to several team programs, and one of the deepest, unfortunately, was to first-year coaching. Our professional first-year Standard/Smooth coach was forced to leave by health issues, and we made the decision to let our professional first-year Latin/Rhythm coach go as well. In their place, we hired four local amateur coaches, one for each style.

As we look forward to next year, the Board, naturally, is interested in finding out whether the experiment worked — did this year's Rookie class suffer in their competitive results due to amateur coaching? Which coaches got

the best results, and which should we hire next year?

Previously, the Board had no way to answer such questions with hard data, and was forced to generate impressions of this year's comparative success by intuition and guesswork. But this visualization makes it relatively to answer the question by laying graphs next to each other.

Before I began, I asked the Board what they expected to find: they believed that Latin style had improved, Standard style had declined, and Rhythm and Smooth styles had remained about the same.

To test this hypothesis, I chose a well-attended competition placed late in the year, Harvard Invitational (March '13 / April '14), which the team had attended in both 2013 and 2014. By graphing results (across three dances) for both years, I could visualize the difference in competitive performance in the same competition, between the two years, for a single style. And, by doing it four times and compiling the results, I could try to answer the general question "Are we better- or worse-off than last year?"

The visualizations are reproduced below. (The top row of panes is HI '13 data; the bottom row is HI '14. Within a pane, the upper portion represents the Harvard team, and more bars stretching farther to the right is better, as is greener bars overall. The lower portion is the rest of the field, for comparison.)

Latin Style: Cha Cha, Rumba, Samba; HI ‘13 vs. HI ‘14

Rhythm Style: Cha Cha, Rumba, Swing; HI ‘13 vs. HI ‘14

Standard Style: Waltz, Quickstep, Foxtrot; HI ‘13 vs. HI ‘14

Smooth Style: Waltz, Tango, Foxtrot; HI ‘13 vs. HI ‘14

Laid out this way, we can see that the Board's predictions are largely incorrect: Standard style did indeed see a decline in success (which was expected), but so did Latin (which was not). Rhythm saw a rather large drop in success (rather than no change). In fact, the only style which saw improvement was Smooth (where no change was predicted), which actually saw quite a large improvement.

Based in part on this data, the Board has made a preliminary decision to hire this year's first-year Smooth coach back next year to teach Rookie Smooth… and second-year Standard. Hopefully, the Board thinks, he'll be able to help this year's rookie class rehabilitate their Standard skills.

B. Further Improvements

While the multi-pane layout is immensely useful for comparing several events against each other, it is slightly inconvenient to have to set up each pane individually, selecting first the competition, and then the event. Since many questions, like the one above, involve comparing the same (or nearly the same) events across different competitions, years, or skill levels, it would be useful to have a way to standardize queries like "display all Standard-style dances for Harvard Invitational 2013" which can be laid out next to (or above) "display all Standard-style dances for Harvard Invitational 2014" in a manner more intuitive than setting up each pane manually.

However, the current design technically has the capacity to lay out basically any combination of events the user desires, and the promising incremental improvements are largely a matter of simplifying the matter of setting up panes in groups, rather than individually.

Of note is the proposed improvement (discussed both in the initial proposal and in the first milestone design lab) of an area-graph view for visualizing the team's performance in the whole, rather than as individuals. After some cursory experimentation, it appears that such a view is unnecessary, or at least redundant. As demonstrated above, the competitors-as-bars view can readily serve to display results from at least six to eight different different events on one screen, which is approximately the practical limit of area graphs as well. Furthermore, variations in color provide extra, useful visual data on performance that would be lost in a simple area graph.

Of course, radically new views might prove useful for answering very different queries ("How does our team compete with MIT's?"), but that probably lies out of the scope of a one-person, two-month project. Going forward, though, I'd definitely be interested in exploring other options and continuing to develop this tool for the use of HBDT (and other teams, I suppose).

IX. Note: Teamwork

By prior arrangement with Prof. Pfister and Alex Lex, I was granted permission to work alone on this project, owing to my obscure interest and difficulty finding a partner as excited about ballroom dance competitions as I was. Thus I've attached no peer evaluation.