CS 171: Project Proposal

Ross Rheingans-Yoo

I. Background and Motivation

A. Personal Background/Introduction

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.

For this project, however, I plan to apply my newfound skills with D3 to create a few dynamic visualizations that draw from competition results scraped from online sources, with a supergoal of answering a few questions that the HBDT executive board has been asking recently:

- Are the recent changes in coaching working? (Are we getting better?)
- What particular styles/dances are we best/worst in?
- Which schools are doing better than us, and in which styles/dances?
- 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 others 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.

B. Previous Research

While online reporting of competition results is relatively new in collegiate ballroom dance, 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), which, nonetheless, will still represent a few thousand judging marks to be wrangled and visualized.

C. 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!

[1] http://www.o2cm.com/
[2] https://github.com/CloudCray/DanceMarkScraper
[3] http://i.imgur.com/gJBUpJJ.jpg

II. Project Objectives

As previously discussed, HBDT is interested in investigating how our competitive results vary between dances, levels, styles, and years. In particular, I'm trying to use this visualization to answer the primary question: "Comparing group X to group Y, who was more competitive?" (in terms of advanced-to-rounds attrition-curve; see "Visualization" below for detail). In particular, the relation between group X and group Y may be:

- same dancers/styles, different events/years
- same events/level/style, different years
- same event, different styles/dances/levels
- same event, different teams

These all answer different questions, and may or may not require subtly different visualization schema (again, see "Visualization" below for detail). In general, though, they're informed by my general experience working with the HBDT executive board in discussions about training and competitions, and follow-up discussions with our team captain and coaches.

III. Data

As noted above, I'll be using an open-source webscraper released by Cloud Cray for scraping competition data from O2CM, a popular results-reporting website. (Relevant links are in earlier footnotes.) I'll be using his Python scripts to perform my own scrapes of my own accesses of O2CM. If I run into excessive difficulty in cleaning the results, I plan to ask Mr. Cray himself for help. (Did I mention that he's a coach for HBDT?)

A note on O2CM: The popular website collates approximately half of the collegiate ballroom dance competitions in the US, and approximately two-thirds of the competitions at which Harvard has competed in the past three years. A competing platform, *zDanceConcepts*, covers approximately all other competitions (and all but one of the competitions in which the Harvard team has competed), but is significantly harder to scrape. In particular, Mr. Cray has not released his *zDanceConcepts* scraping code, claiming that they are not yet robust enough for widespread use, and that the results typically require excessive cleaning in order to be usable.

IV. Data Processing

As noted, I don't expect to have to do much manual data cleaning, but will have to do a lot of aggregation and processing. From a list of individual judges' marks[4], I will extract two sorts of information for each couple, per event entered, per dance:

- for final rounds, the placement recommended by each judge
- for all other rounds, which judges recommended advancement to the next round

These will allow me to aggregate further statistics for each couple, including:

- the last round reached
- for each couple who didn't make some round, how many judges recommended advancement (and how many would have been necessary to be advanced)
- for each finalist couple, their overall place
- for each non-finalist couple, their approximate overall place

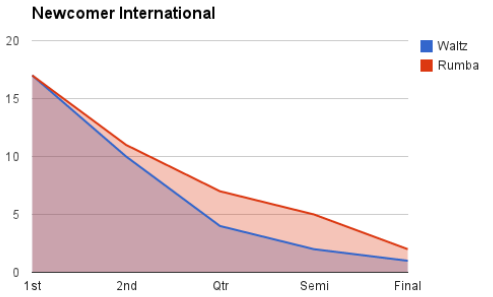
and, for a given subset of dancers (in particular, one generated by school affiliation), the number advanced to a given round in a given style/level/dance.

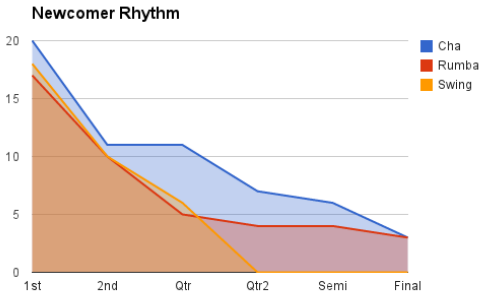
I expect to perform most or all of this data processing in Python, exporting csv or json files to be read by D3-based web scripts.

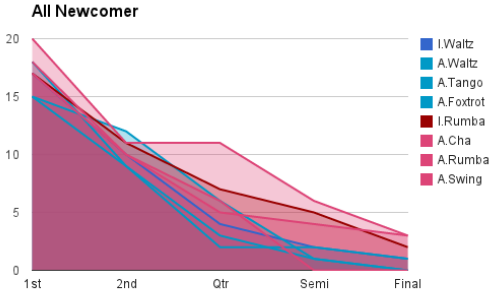
V. Visualization

In attempting to answer the question "How is our team doing?", I compiled the following graphs based on results at one competition last semester:

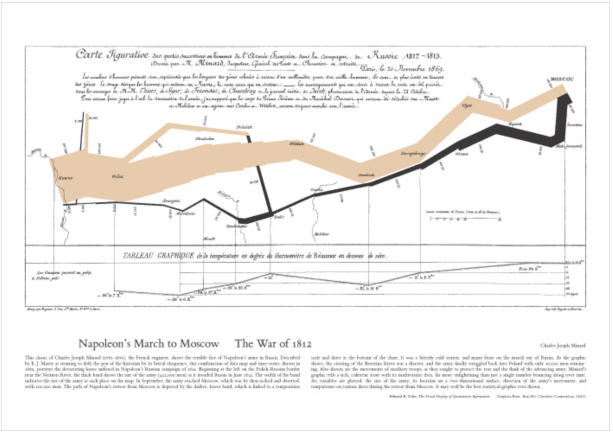
[4]Example: https://github.com/CloudCray/DanceMarkScraper/blob/master/tests/samples/output_data.csv







They display the number of Harvard dancers (y-axis) advancing to each round (x-axis) in several dances (in the first, two International style dances; the second, three in American Rhythm style; the third, all dances at the competition) at the Newcomer level. (They violate several design principles that we've discussed in class, but keep in mind that I generated them—using Google Docs—long before the first day of 171…) Of course, I owe some debt of inspiration to the famous graphic illustrating Napoleon's historically disastrous march into Russia:



My finished project would likely include a cleaned-up, interactive version of my rudimentary "attrition-graph" visualizations (for comparing team performance across different dances/styles), along with other similar graph-based visualizations for comparing competing schools (with a stacked area graph) or individual dancers/groups across years (side-by-side, overlaid, or using other methods).

VI. Must-Have Features

In order to adequately address the "Project Objectives" laid out above, I expect that my project will have to be able to:

- generate an attrition-graph for a single dance/level/style/event, for a given set of dancers.
- generate multiple (at least two, but preferably a variable number) of graphs side-by-side.
- generate a 'stacked' version of such a graph, which plots two or more sets of dancers against each other.

VII. Optional Features

Additional features which would be nice, but by no means necessary include:

- interactivity, to reveal details on an individual couple's results (and cross-compare them across multiple graphs, if appropriate).
- some intuitive interface for identifying and browsing relevant, interesting comparisions (of the sort described earlier).

It's possible that other features will present themselves in the course of development, but for now I'd prefer to make those decisions based on early prototypes and experimentation, rather than mere design sketches and speculation.

VIII. Project Schedule

Realistically speaking, I won't have much time to devote to this project until we strike *Les Phys*, the musical I'm on-staff for—that'll be April 6. Up until that point, I may try to find some time to get basic templates and functionality online, but I honestly don't expect much significant progress to take place. In the next week, then I'll work toward the first functional prototype, a system which takes webscraped results from a given competition, processes and aggregates them, and displays an attrition-graph for a given event/set of dancers. This prototype will be targeted for April 9.

The week of April 14 is scheduled for project review with TFs—I plan to spend the previous weekend working to begin extending the prototype to include the full "Must-Have Features", to have a good idea of the design challenges that will lie ahead of me. Tentatively speaking, I hope to have these must-haves completed by May 24, with an absolute deadline of May 27. This leaves me between a week and four days to finalize the project, its documentation, and presentation, including screencast.