Appendices

A Sampling orderings uniformly at random conditioned on a phylogeny

In general, the subtree rooted at vertex x defines a partial ordering on its own mentions. To sample a total ordering $\boldsymbol{i}_x$ uniformly at random from among those compatible with that partial ordering, first recursively sample M orderings $\boldsymbol{i}_{y_1}, \ldots, \boldsymbol{i}_{y_M}$ compatible with the M subtrees rooted at x's children. Then uniformly sample an interleaving of the M orderings,and prepend x itself to this interleaving to obtain $\boldsymbol{i}_x$. To sample an interleaving, select one of the input orderings $\boldsymbol{i}_y$ at random, with probability proportional to its size $|\boldsymbol{i}_y|$, and print and delete its first element. Repeating this step until all of the input orderings are empty will print a random interleaving. Note that in the base case where x is a leaf (so $M = 0$), this procedure terminates immediately, having printed the empty ordering. Our $\boldsymbol{i}_\diamondsuit$ is the output of running this recursive process with $x = \diamondsuit$.

B Twitter Grammy corpus

B.1 Collection

Using the Twitter 1% streaming API, we collected all tweets during the 2013 Grammy music awards ceremony, which occurred on Feb 10, 2013 between 8pm eastern (1:00am GMT) and 11:30pm (4:30 GMT). We used Carmen geolocation (Dredze et al., 2013) to identify tweets that originated in the United States or Canada and removed tweets that did not have a language of English selected as the UI for the tweet author. This yielded a total of 564,892 tweets. We then selected tweets that contained the string "grammy" (case insensitive), reducing the set to 50,429 tweets. These tweets were processed for POS and NER using the University of Washington Twitter NLP tools [13] (Ritter et al., 2011). Tweets that did not include a person mention were removed. For simplicity, we selected a single person reference per tweet. The final set contained 15,736 tweets. Of these, 5000 have been annotated for entities.

B.2 Annotation

A first human annotator made a first pass of 1,000 tweets and then considered the remaining 4,000 tweets. This provided an opportunity to refine the annotation guidelines after reviewing some of the data. The annotator was asked to assign a unique integer to each entity and to annotate each tweet containing a mention of that person with the corresponding integer. Additionally, the annotator was asked to fix incorrect mention strings. If the extracted mention was incorrect or referred to a non-person, it was removed. If it was mostly correct, but omitted/excluded a token, the annotator corrected it. Similar to Guo et al. (2013), ambiguous mentions were removed. However, unlike their annotation effort, all persons, including those not in Wikipedia, were included. Mentions that were comprised of usernames were excluded (e.g. @taylorswift13). Following this protocol, the annotator removed 423 tweets. A second annotator inspected the annotations to correct mistakes and fix ambiguous references. The final annotated corpus contains 4,577 annotated tweets and 273 distinct entities. This corpus was then split into five folds by first sorting the entities by number of mentions, then performing systematic sampling of the entities on the sorted list.

[13] https://github.com/aritter/twitter_nlp