A Maximum Likelihood Approach For Selecting Sets of Alternatives

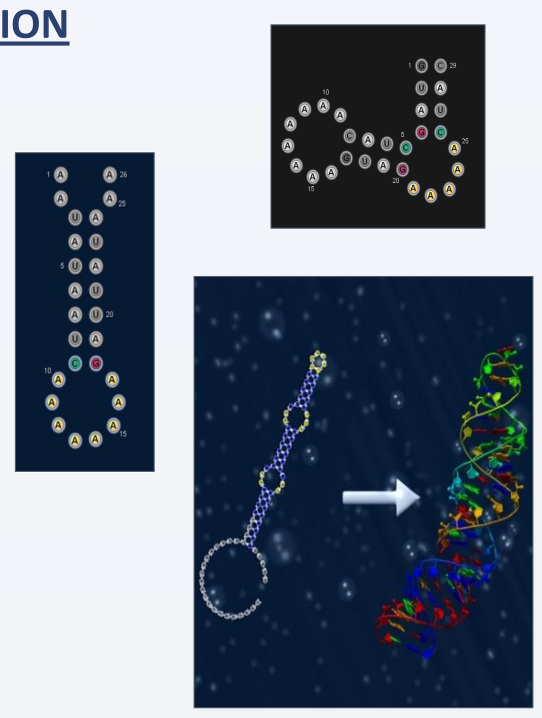
Ariel D. Procaccia and Sashank J. Reddi and Nisarg Shah

School of Computer Science, Carnegie Mellon University.

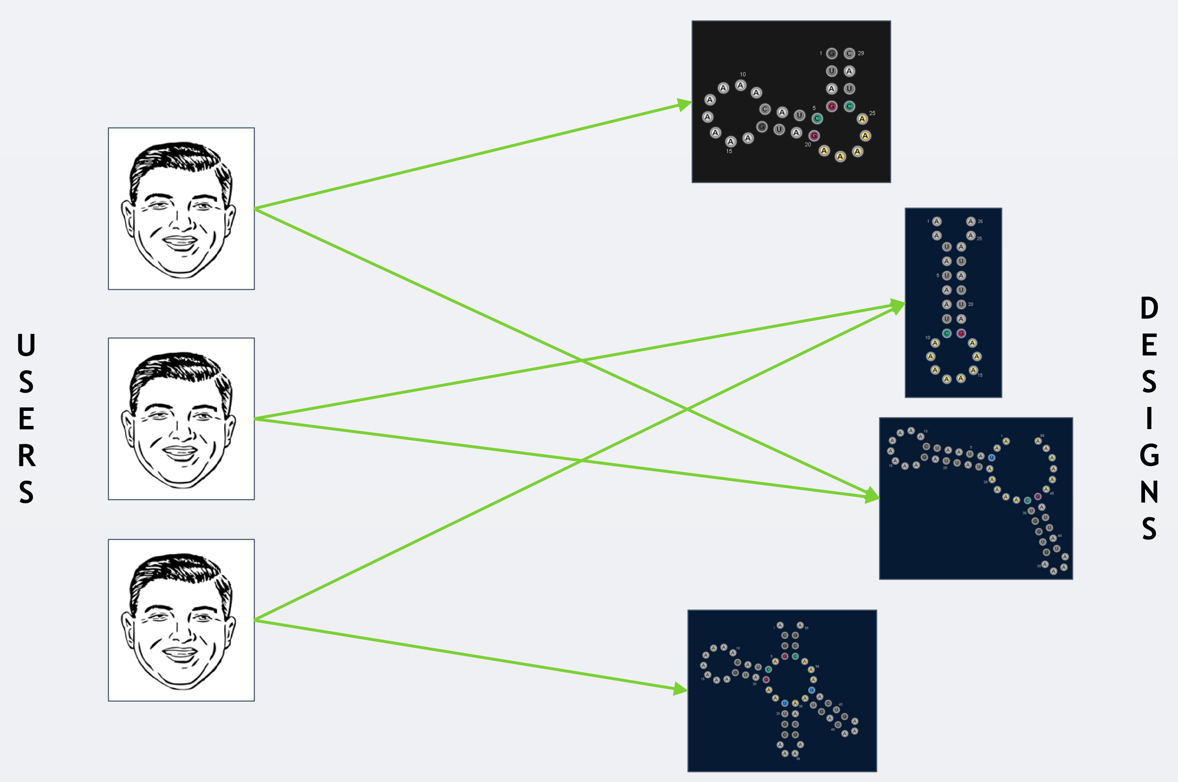
MOTIVATION

EteRNA

- A game on RNA folding.
- RNA folds into stable shapes.
- Need to generate RNA designs which fold into a required shape.
- Can ultimately lead to ground-breaking discoveries such as an RNA random access memory, a nanoLED, a switch that detects cancer as soon as it starts spreading etc.
- Difficulty: Hard to predict how well a particular RNA design will fold in reality. Thus needs to be synthesized in the lab.



- Tens of thousands of participants propose their RNA designs for a particular goal.
- We want to find out which design folds best in reality. But we have the budget to synthesize only 8 designs in the lab.
- Users vote for designs and help pick out the 8 designs such that the best design is very likely to be one of them, which will be singled out after synthesis.
- Currently, the designs are picked using a rule known as *k-approval* where each user picks out the 8 designs that he thinks are the best for this purpose. The 8 designs which get picked out by most users are synthesized in the lab.



BUT IS THIS THE BEST WAY TO CHOOSE THE 8 DESIGNS TO BE SYNTHESIZED? THE ANSWER IS…..NO !

PROBLEM STATEMENT

Data:

- 'm' alternatives (RNA designs in the above example).
- Underlying (unknown) true ranking/order among the alternatives 'σ^*' (not subjective)
- 'n' "preferences" among the alternatives or "votes" sampled from a distribution 'Γ' centered around σ^*. Need not be total orders among the alternatives.
- The number of alternatives to be selected - 'k'.

Objective:

- Find a target (according to some criteria) subset or tuple of alternatives of size k.
- Since σ^* is unknown, use maximum likelihood estimation given the n votes from the distribution Γ.

DISTRIBUTIONSOF VOTES

MODEL 1: Noisy Comparisons

- n preferences between each pair of alternatives (a,b).
- 'a $\succ$ b' with probability 'p' and 'b $\succ$ a' with probability 1-p if a $\succ$ b in σ^* and vice-versa.

MODEL 2: Noisy Orders *(Also, Mallow's Model / Condorcet Noise Model)*

- n total orders among the alternatives.
- Probability of an order decreases exponentially as the distance from σ^* increases.
- Pairwise Distance (Also, Kendall's Tau Distance): $d_K(\sigma,\sigma^*) = \text{number of disagreements between } \sigma \text{ and } \sigma^* \text{on pairs of alternatives}$

Our Generalization: Noisy Choice Model

- Any preference dataset D and its distribution Γ that satisfy the following properties.
- Need not be total orders or pairwise preferences.
- n_{ab} = # (a $\succ$ b) in D. This should be clear from the preference format.

Properties:

1) $n_{ab} + n_{ba} = n, \text{for every } a \neq b.$

2) $\Pr[D|\sigma^*] = \frac{\gamma^{d_K(\sigma^*,D)}}{Z_\gamma}$, where $d_K(\sigma^*,D) = \sum_{(a,b)|a \succ b \text{ in } \sigma^*} n_{ba}$.

- γ is the noise level. $\gamma \to 0$ represents completely noise free distribution and $\gamma = 1$ represents the uniform distribution.

RESULTS I: INCLUDE-TOP

Objective:

- Select a subset of size k that maximizes the chances of including the best alternative (which is at the top in the true order σ^*).

Applications:

- New Product Development (NPD) - shortlisting a few designs through a market survey.
- Cloud Computing - choosing nodes for speculative execution.
- Crowdsourcing.

Why simple majority does not work!

a > b > c > … > y > z
a > b > c > … > y > z
x > b > c > … > z > a
x > b > c > … > z > a
y > b > c > … > z > a

k = 2 -- 'a' and 'x' are selected.
'b' is more likely to be the best!

How to select the optimal k-subset of candidates?

- ***Theorem:*** INCLUDE-TOP is ***NP-hard*** for both noisy orders and noisy comparisons (and hence for noisy choice model) for any $1 \leq k \leq m-1$.

- The hardness result uses the case when $\gamma \approx 0$, i.e., when there is little noise. But in that case, we don't need the best because any reasonable rule would work well, if not optimally.
- For the other extreme when there is high noise, it turns out that we can indeed perform optimally in polynomial time.

- ***Extended Scoring Method:*** $SC(a) = \sum_{b \neq a} n_{ab}$ (number of pairwise wins).
- ***Theorem:*** When there is high noise (γ is sufficiently large), top k candidates according to ***ESM*** maximize the chances of including the best alternative.
- For noisy orders, reduces to a famous rule known as the *Borda count*.

- ***Q:*** Okay, we included the best. What about the other alternatives chosen? Are they very bad? Intuition says NO. If they were likely to be the best, they must be good themselves. Turns out, that is exactly right!

RESULTS II: TOP-SUBSET

Objective:

- Select a subset of size k that maximizes the chances that it is the subset of top k candidates in the true order σ^*.

- Appears in practice, for example, in team building.
- Also ***NP-hard*** for both noisy orders and noisy comparisons for any $1 \leq k \leq m-1$.
- When there is high noise (γ is sufficiently large), top k candidates according to ***ESM*** also form the optimal k-subset for the TOP-SUBSET objective.

Implication:

- Top k candidates in the ESM do not only ensure that the top candidate is included with high probability, but also ensure that the other candidates are very likely to be the next k-1 best candidates.

RESULTS III: TOP-TUPLE

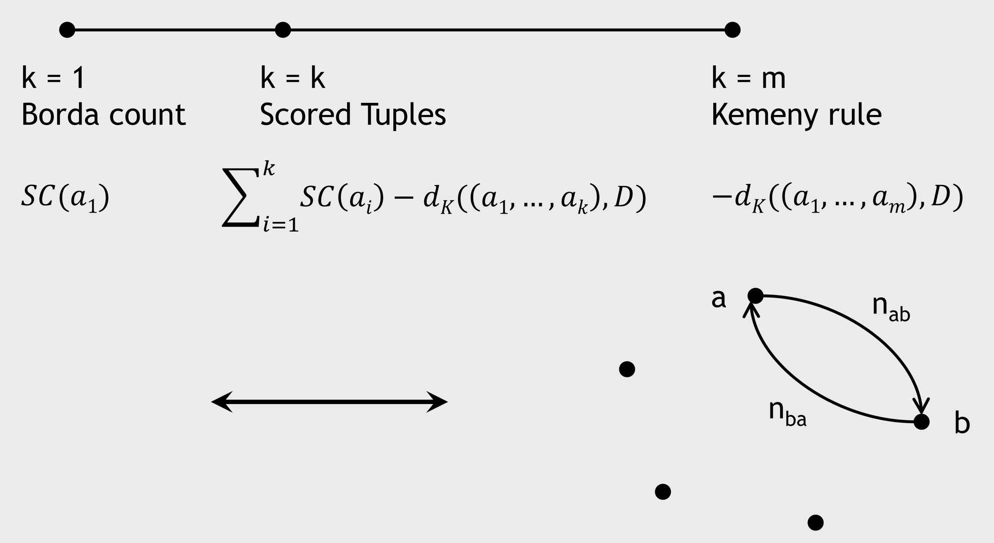
Objective:

- Select an ordered tuple of size k that maximizes the chances that it is the k-prefix of the true order σ^*.
- ***Q: Why is it not the same as TOP-SUBSET?*** Sum of likelihoods of various orders…
- Appears in practice, for example, in selecting a committee (e.g., president, vice-president etc.) and in combining search results from different search engines.

- As before, the objective is ***NP-hard*** for $1 \leq k \leq m$ and is tractable when the noise is sufficiently high.
- For k=1, the objective reduces to finding the candidate which is most likely to be the best candidate and for k=m, it reduces to finding an MLE for the true order.
- Thus for noisy orders with high noise, it reduces to Borda count for k=1 and Kemeny rule for k=m. In fact, the optimal rules for various values of k in this case form a continuum between the two voting rules.

Scored Tuples Method:

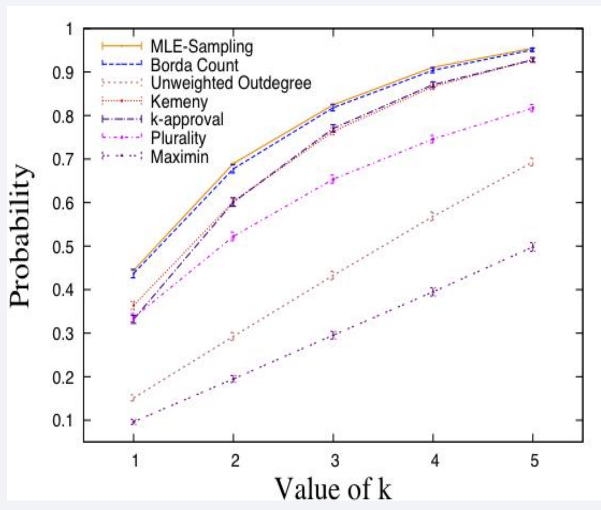
- $SC(a_1, a_2, \dots, a_k) = \sum_{i=1}^{k} SC(a_i) - d_K((a_1, a_2, \dots, a_k), D)$
- Choose the k-tuple that maximizes the score.



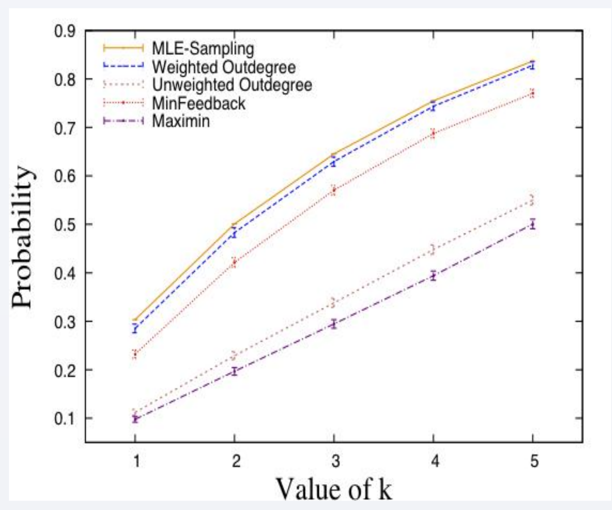
Pairwise Candidate Graph

SIMULATIONS

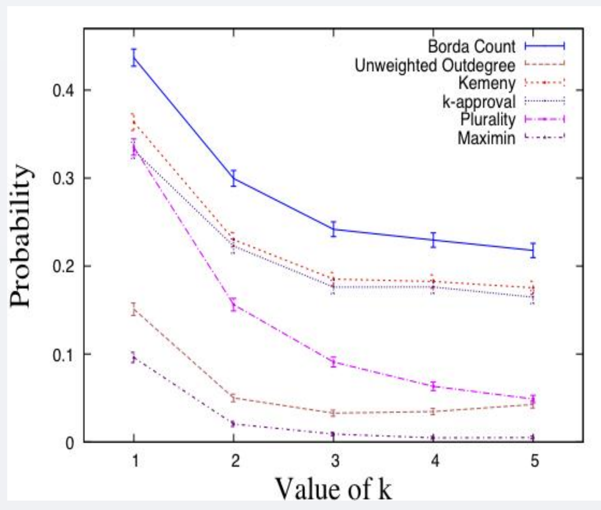
- Setting: m=10, n=10, p=0.55 ($\gamma \approx 0.818$)
- Same results are observed for other values of m, n and γ as well.



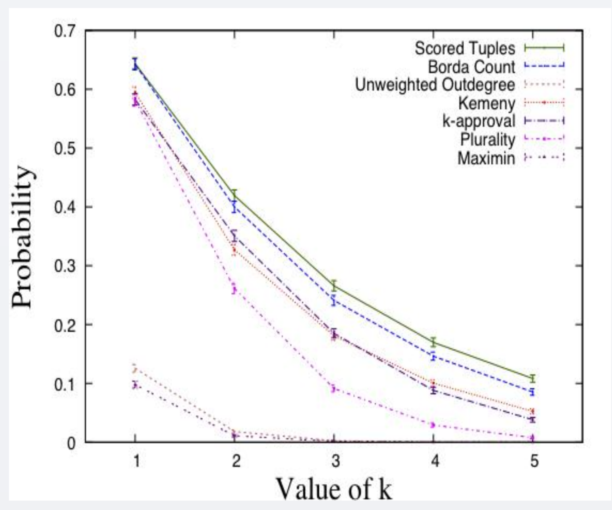
INCLUDE-TOP : Noisy Orders



INCLUDE-TOP : Noisy Comparisons



TOP-SUBSET: Noisy Orders



TOP-TUPLE : Noisy Orders

MAIN CONTRIBUTION

Noisy Choice Model

- A unifying model that captures several well studied preference types and their distributions from the exponential family.
- Generalizes noisy orders, noisy comparisons, noisy partial orders of a fixed length etc.
- The distributions in noisy choice model share the same optimal method for our objectives.

- ***Open Question:*** What other properties do these distributions share?

Ordinal objectives for selecting sets of alternatives

- Several practical cases where more than one winner needs to be selected.
- INCLUDE-TOP, TOP-SUBSET and TOP-TUPLE serve as useful objectives when cardinal comparison of the alternatives is not possible.

- ***Research Direction:*** INCLUDE-TOP and TOP-SUBSET share the same optimal method in this case. This hints at a possibility of an intrinsic connection between the two objectives, hopefully in a more general setting.

Connection between Borda count and Kemeny ranking

- Connected Borda count and Kemeny ranking via a continuum of 'optimal' voting rules.
- Connection has elementary interpretation in the pairwise candidate graph.
- This might be of independent interest to social choice theory.

EXTENSIONS

'High Likelihood' instead of 'Maximum Likelihood'

- Instead of maximizing the probability of including the top candidate, just ensure that the probability is sufficiently high.
- More candidates need to be included in the subset to increase the probability.

- ***Question:*** How many candidates need to be chosen?
  - Preliminary results available for Extended Scoring Method.

Solving the objectives with high probability

- Selecting the optimal k-subset for INCLUDE-TOP is ***NP-hard*** but it can be computed with high probability.
- This approach relies on sampling from an extension of Mallow's model known as the Generalized Mallow's Model.

- ***Open Question:*** Can sampling from the Generalized Mallow's Model can be done in polynomial time?

Data Allocation

- Current method: input dataset is assumed to be given and a subset of alternatives is selected optimizing certain objective functions.
- Consider an extension where there are multiple users available with different precision level and there is a limit on the number of evaluations that each user can perform.
- Two processes: allocation (static or dynamic) of alternatives to the users to get the dataset and combining these evaluations to compute the answer.

- ***Open Question:*** How to design a mechanism that performs both functions in a way that jointly optimizes the objective function?