Dear Jens,

Me and Ahmed have been working exhaustively on the project over the christmas holidays. So for we have been able to create a random HOT. But mostly our work has been trying to understand the theory by reading all the references for your article (we have read around 10 papers in addition to other sources to understand).

Currently we are wondering if the following approach is the correct one. Our approach is divided in three steps, in summary.

1. Create a random HOT using synthetic data.
2. Create datapoints using the HOT constructed in step 1.
3. Re-create a HOT using the data 1 step 2 and compare it to the original HOT created in step 1.

Is this approach correct? If that is the case, then. Is it correct that the datapoints created in step two are sets of mutations (and only that), where the overall frequency of the mutations of the union of the sets corresponds to the probabilities in the original HOT? For example. {xp, 13p, -qr}{xp, +qr, p} … {x, +1p, -Xr} If so which probability distributions we are sure that Px are evident from the sets but what about Pz.

Please check our steps in detail below because we are really lost and it is taking us much time trying to understand the algorithm. We are recursively reading papers and documents.

1. Initially we are uniformly choosing

*pz*(*u*) = Pr[*Z*(*u*) = 1*|Z*(*p*(*u*)) = 1]

*px*(*u*) = Pr[*X*(*u*) = 1*|Z*(*u*) = 1]

*ez*(*u*) = Pr[*Z*(*u*) = 1*|Z*(*p*(*u*)) = 0]

*ex*(*u*) = Pr[*X*(*u*) = 1*|Z*(*u*) = 0]

as the global parameters for the algorithm as follows:

* 1. *pz ∈* [0*.*1*,* 1*.*0]*,*
  2. (1 *− px*)*, ex, ez ∈* [0*.*01*, q*]*; q ∈ {*0*.*05*,* 0*.*10*,* 0*.*25*,* 0*.*50*}*.

1. We will generate random HOTS by variating the following:
   1. Number of vertices: 10, 25, 40 (3 possibilities)
   2. *q ∈ {*0*.*05*,* 0*.*10*,* 0*.*25*,* 0*.*50*}* (4 possibilities)
   3. 100 HOTS per each number of vertices and q (100 possibilities)

The total number of synthetic HOTS generated would be 3x4x100 = 1200

1. We then generate sets of the mutations by traversing the HOTS, the number of sets generated would be 100,500,2000,5000.
2. Each dataset was then passed to the algorithm and the resulting HOT was compared to the original HOT.

If there is any suggestion, please let us know.

Yours,

Erik