Appendix

Anonymous University, Anonymous Address `anonymous@anonymous.com`

1 Appendix

1.1 Filtering Metrics

We evaluate our confidence in several tested metrics as a reliable score for comparisons. Besides the metrics already discussed in Section **??**, we also evaluated the following ones:

Rank Correlation (RankCorr). This metric uses the Kendall's tau rank correlation coefficient [3]. This coefficient is usually employed to evaluate rankings and was used in the evaluation of several methods [1, 2, 5]. It ranges from -1 up to 1, with the maximum value being achieved when ranks are equal. Rank correlation $RankCorr(\boldsymbol{A}_i, \hat{\boldsymbol{A}}_i)$ is measured for each node in the graph. It captures the correlation between causal parameters α_{ba} and the number of events in the ground truth per row. As Figueiredo et al. [2] discussed, the Kendall correlation suffers when recovering sparse matrices, as most cells are zeroes.

Relative Error (RelErr): The average of the relative errors between each cell of the estimated adjacency matrix and the ground truth:

$RelErr(\boldsymbol{A}_i, \hat{\boldsymbol{A}}_i) = \frac{1}{K}\sum_j \frac{|\boldsymbol{A}_{ij} - \hat{\boldsymbol{A}}_{ij}|}{|\boldsymbol{A}_{ij}|}\mathbb{1}_{\boldsymbol{A}_{ij} \neq 0} + |\hat{\boldsymbol{A}}_{ij}|\mathbb{1}_{\boldsymbol{A}_{ij}=0}$.

The Relative Error was also used for the evaluation of most methods [1, 2, 4, 5], but similarly to RankCorr, it is not completely suitable for sparse datasets: note that there is a constant penalization to avoid division by zero, which will result in a metric dominated by this penalization.

Precision at n (P@n) and Average Precision at n (AP@n). Still focused on rankings, we also measure the Precision@n:

$P@n(\boldsymbol{A}_i, \hat{\boldsymbol{A}}_i) = |\mathbb{T}_n(\boldsymbol{A}_i) \cap \mathbb{T}_n(\hat{\boldsymbol{A}}_i)|/n$, used in [2]. Here, $\mathbb{T}_n(\boldsymbol{A}_i)$ are the top n elements in $\boldsymbol{A}_i$ ordered by their value. The $P@n$ metric avoids the problem of sparsity, as it only considers the edges with the highest weight. However, it ignores the distribution of weights for the edges that are not in the top n. Average of Precision at n (AP@n) aggregates the Precision at n for every possible n. By doing so, this metric solves the issue of choosing a specific n. However, equal weight is given to all of these choices.

Area under the curve (AUC). To calculate the area under the ROC curve we define the binary classification task of predicting whether $\boldsymbol{A}_{ij} > 0$ according to the value of $\hat{\boldsymbol{A}}_{ij}$. The true positive and negative rates are calculated for this task and used to calculate the AUC. The AUC allows one to analyze how well the models estimate the unweighted network. On the other hand, the information on the strength of the influence for a given pair of processes is lost. AUC scores range from 0 to 1.

After optimizing each method using the MLE approach we evaluated whether the metric values are meaningful. This was done using a null-hypothesis test to assess whether for each row of the matrix the method provides a statistically significant value.

Results are presented for the hyper-parameter set with best results. The null-hypothesis was tested with the aforementioned permutation test. When this hypothesis is rejected (a p-value of $p < 0.05$), we can state that the metric may not be explained due to random chance (e.g., permutations). When the test is not rejected, the metric may be explained due to chance. It is common for tests to fail (i.e., failure to reject the null-hypothesis) when sample sizes are small. Unfortunately, this is quite common in sparse networks as the ones we explore. The calculation of p-values (p) was computed considering 1000 permutations.

Our permutation test results for the evaluation the metrics was done using six out the seven methods. That is, we excluded NetInf as it only executes on a single dataset. We shall evaluate NetInf further on using a fair setting for the method. In the six remaining methods, we measured each of the seven metrics – P@n, RelErr, RankCorr, AUC, AP@n, NRMSE and NDCG. For each metric, we computed the fraction of node by hyper-parameter choices where the metric led to statistically significant values. Overall, we found that only for AP@n (fraction of 0.51), NRMSE (0.53) and NDCG (0.55), the fraction of nodes (or rows) with $p < 0.05$ was above 0.5. Statistical significance on all rows is very unlikely due to the sparsity of the matrices (e.g., degrees follow a power-law data thus in some rows we have very few positive values). However, for all other metrics this fraction was much lower being below 0.3 in all settings. This result indicates that the other four (P@n, RelErr, RankCorr and AUC) scores are more unsuitable than the former three. That is, they have more results explained due to chance. Figure 1 shows the p-values distribution for each metric.

From this first analysis, we can argue that by testing on at least six methods, only three metrics AP@n, NRMSE and NDCG are deemed reliable. It is interesting that two of these metrics focus on rankings (AP@n and NDCG), while one focuses on the difference between estimated and ground truth matrices (NRMSE). Based on these results and due to space constraints, in the we present our findings for AP@n, NRMSE and NDCG.

1.2 Full results for the cascades method comparison

Method/Metric	**NRMSE**	**AUC**	**RelErr**
HkEM	0.139 (0.138, 0.143)	0.694 (0.670, 0.702)	553.8 (503.6, 782.7)
ADM4	**0.138 (0.137, 0.141)**	0.732 (0.691, 0.737)	92.7 (68.4, 139.8)
HC	0.141 (0.141, 0.143)	0.692 (0.642, 0.690)	**39.2 (31.8, 53.7)**
GB	0.147 (0.142, 0.151)	0.674 (0.626, 0.677)	383.5 (198.0, 748.6)
ExpKern	0.616 (0.590, 0.654)	0.709 (0.681, 0.713)	16095.6 (13716.7, 18592.1)
NetInf	0.398	**0.737**	10677.9
NullModel	0.205 (0.196, 0.210)	0.500 (0.482, 0.523)	889.67 (101.0, 5313.8)

Table 1. NRMSE, AUC and RelErr for the Memetracker-cascades dataset.

1.3 Code for reproducibility

https://www.dropbox.com/sh/nm1a183z7l951yj/AABJQI-jc2IFzpsVEeLBRKa3a?dl=0

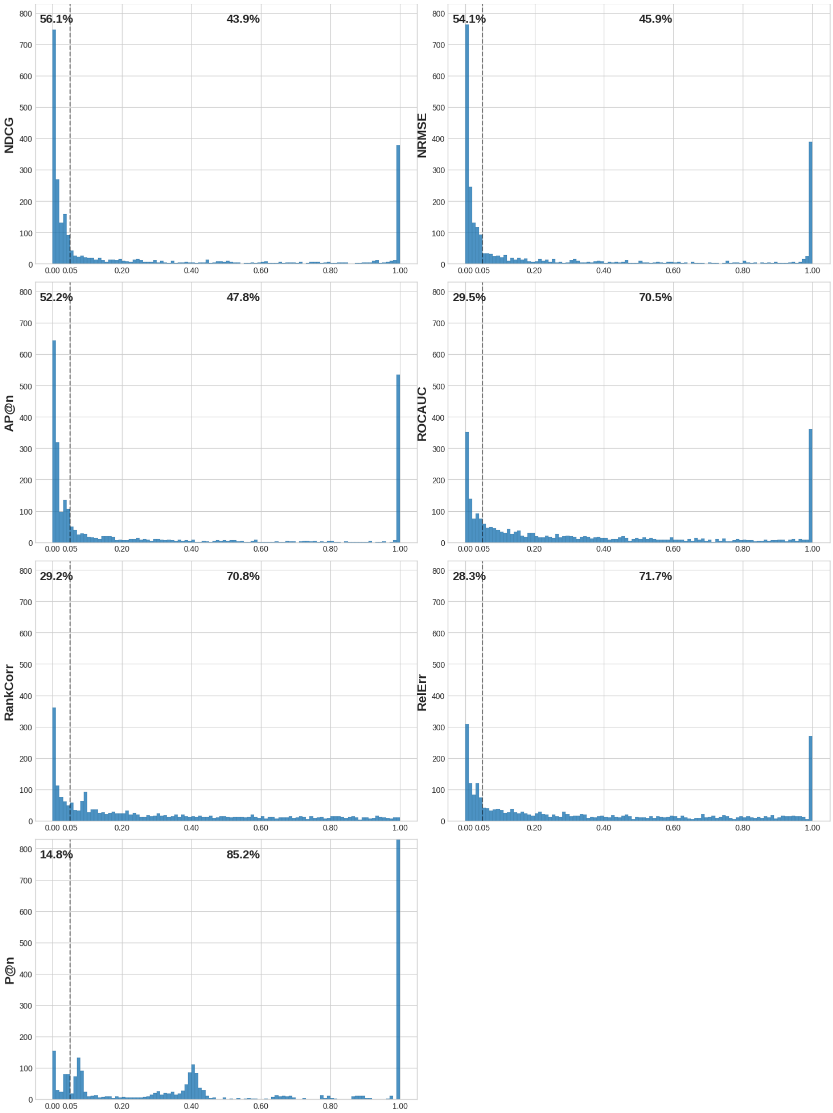
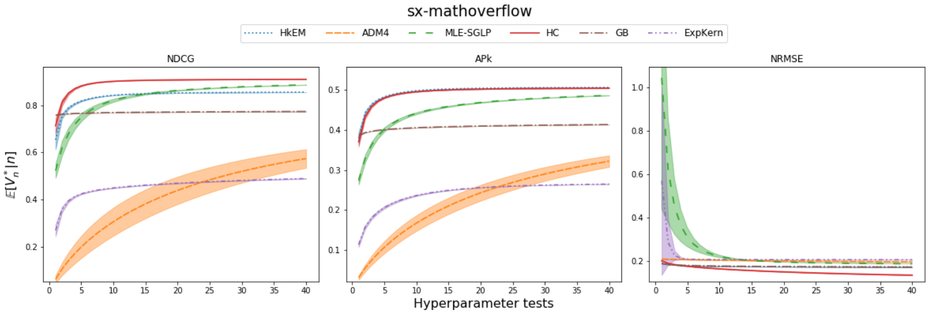
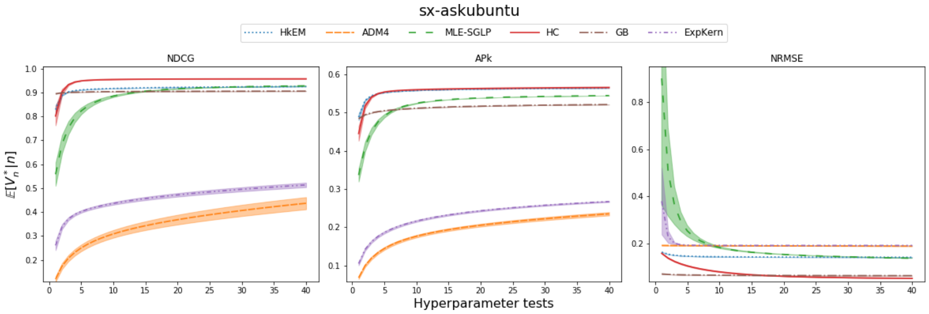
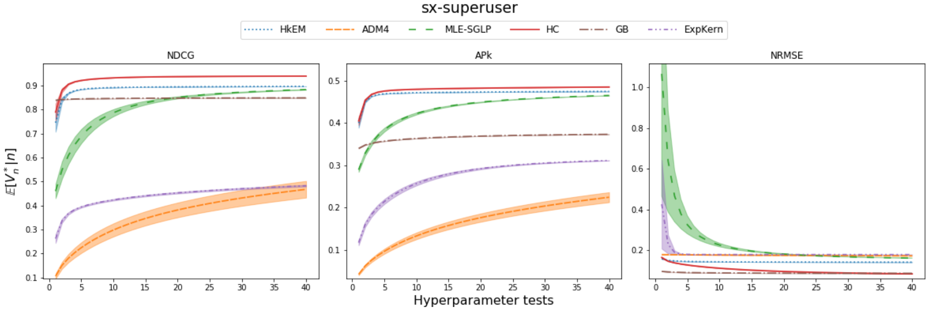


Fig. 1. Histogram for the p-value results of the hypothesis test, aggregated by metric. The left percentages are how often $p < 0.05$.

1.4 Additional plots for the expected value of the best score







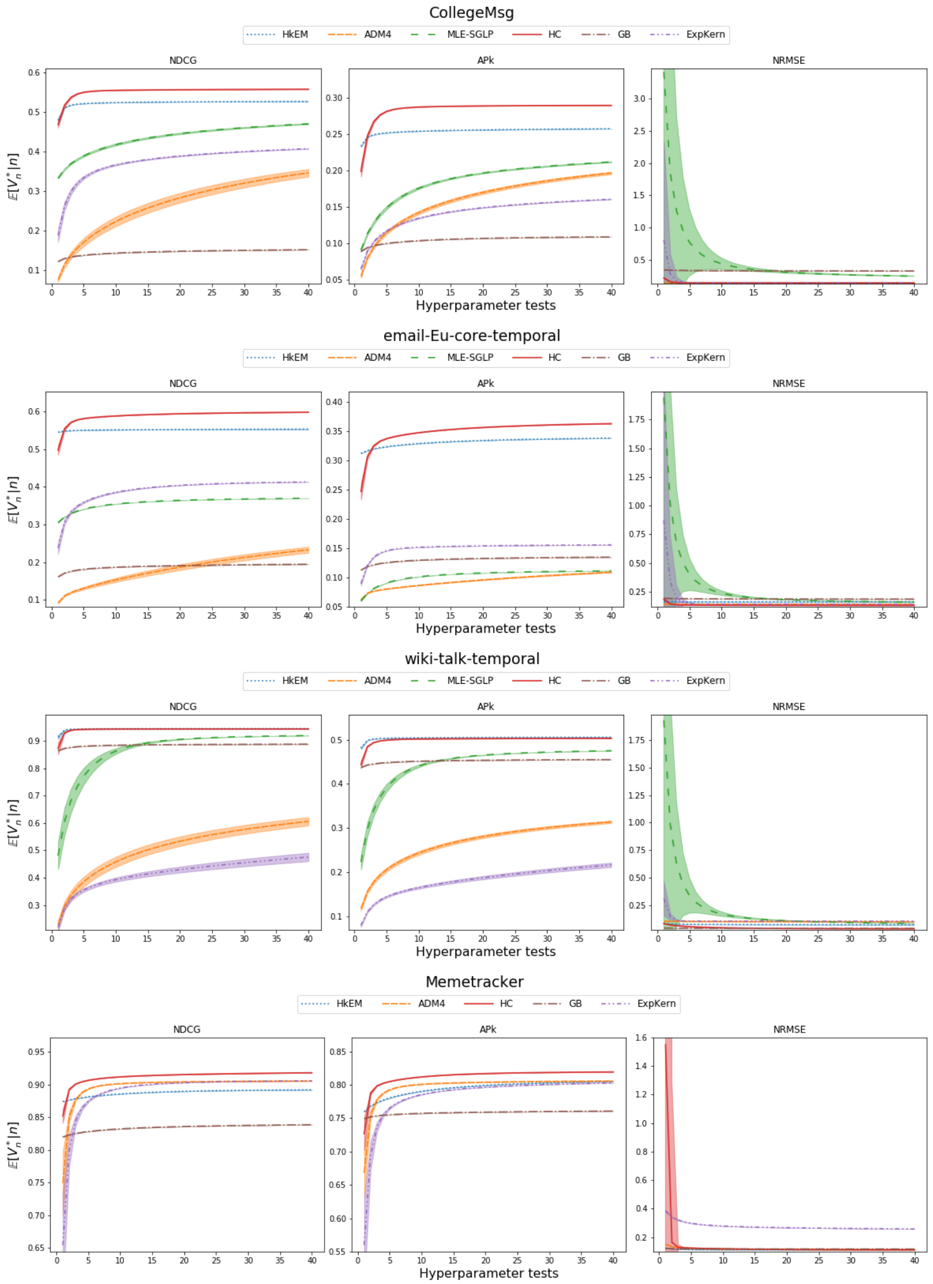
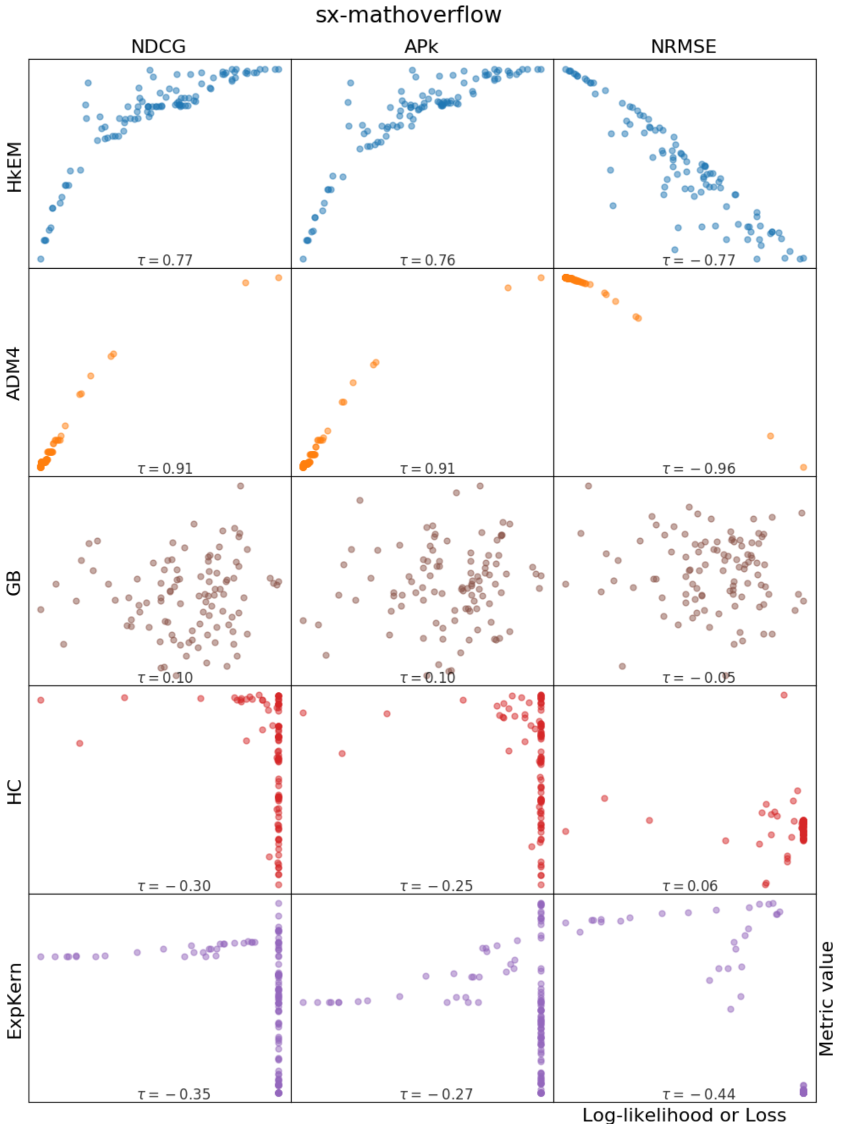
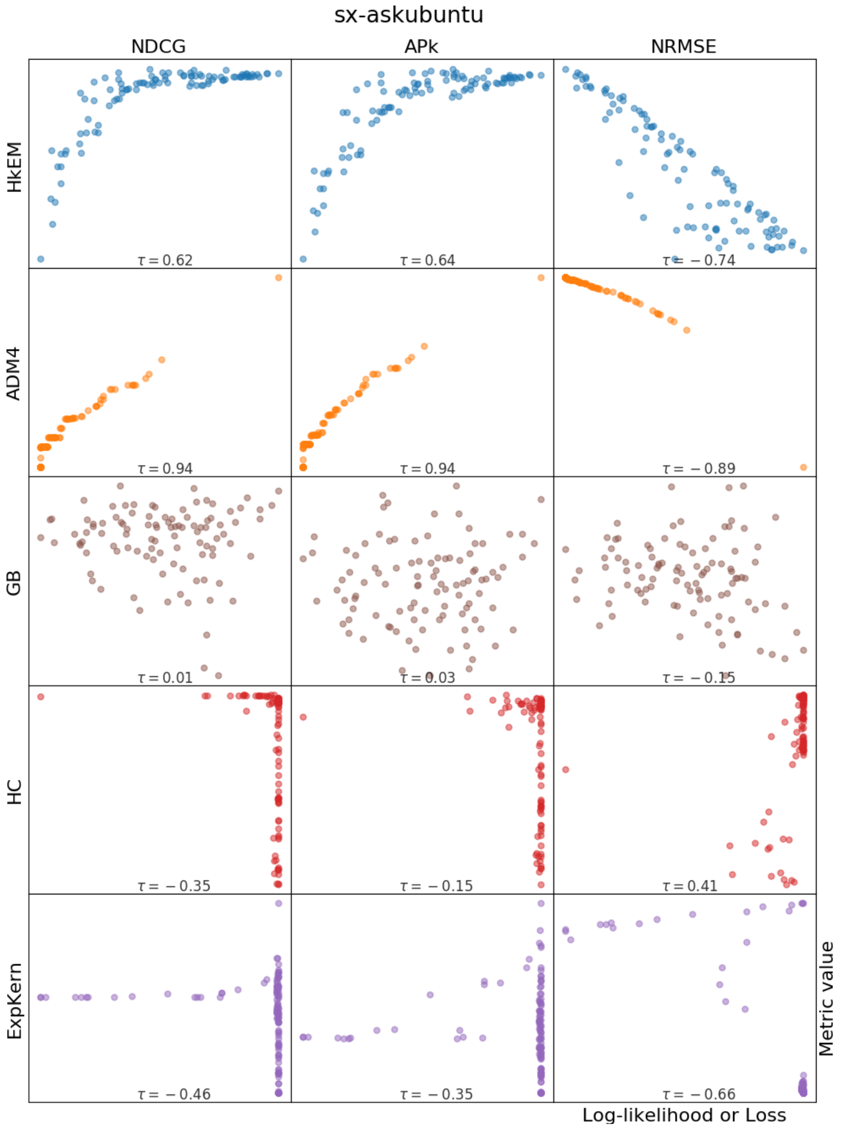
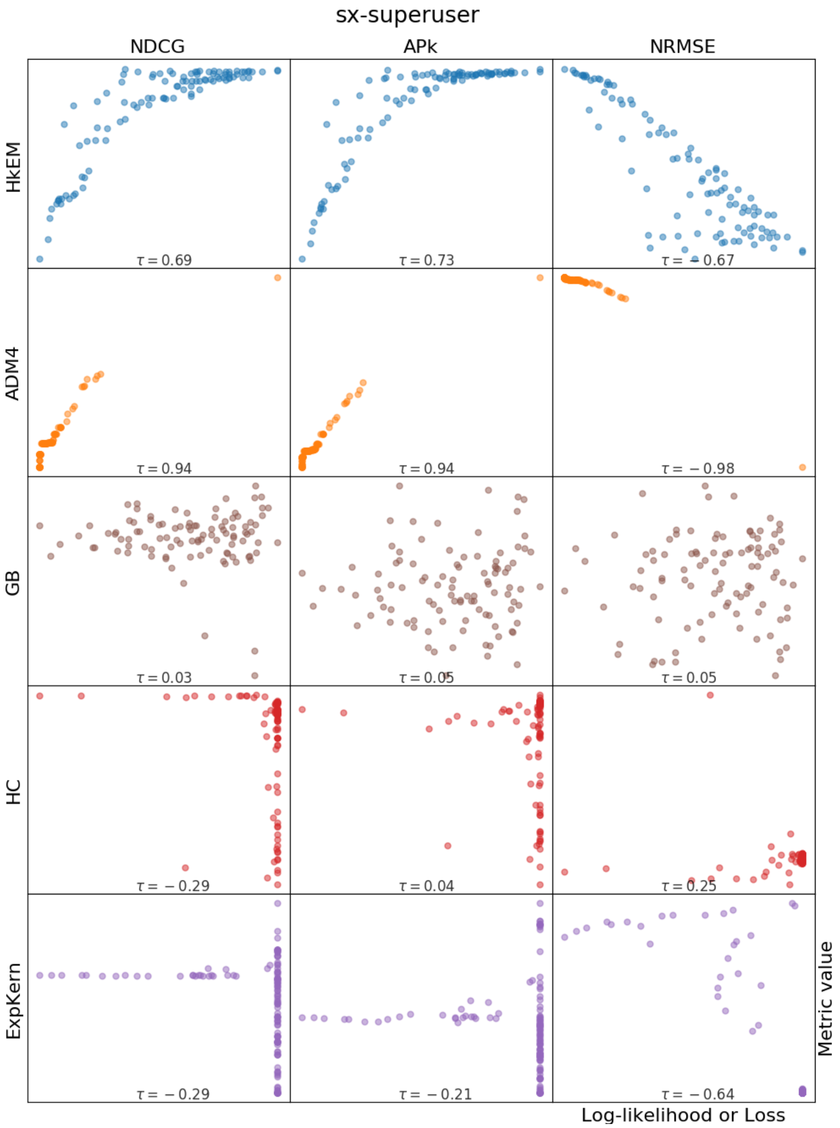


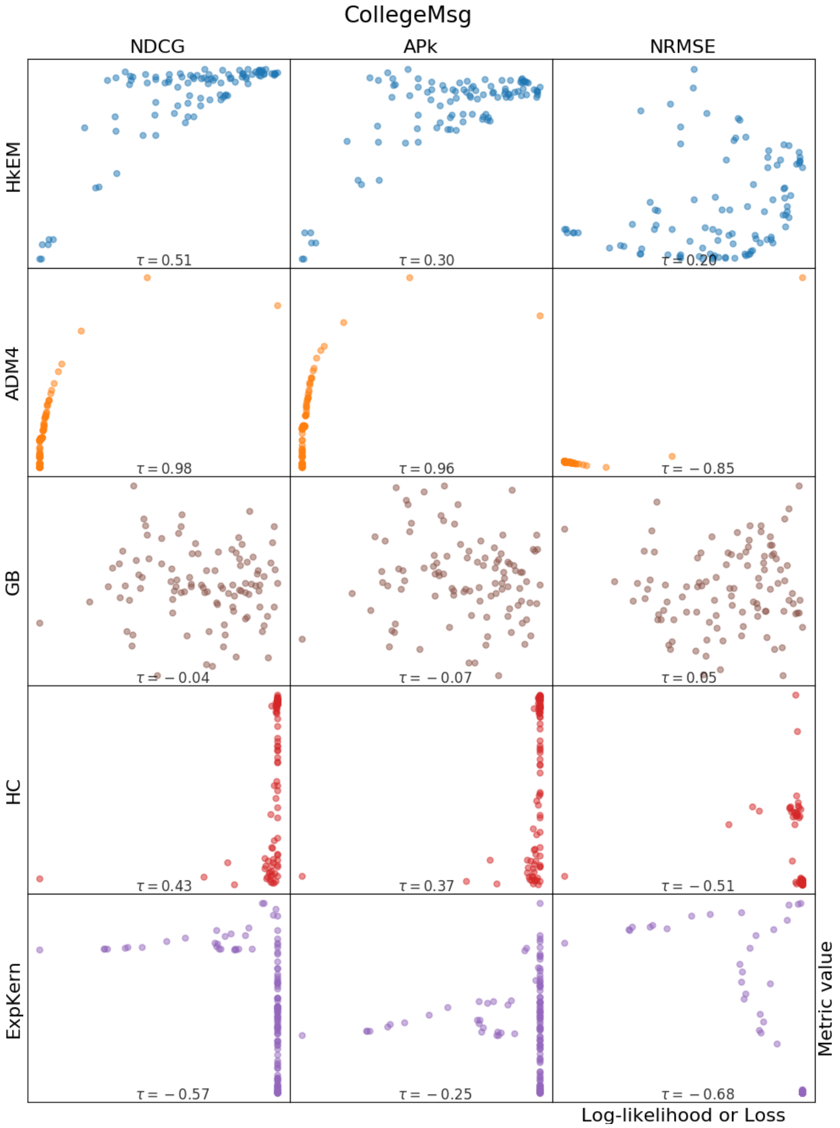
Fig. 2. Expected value of the best score according to the number of hyper-parameter sets tested for all datasets.

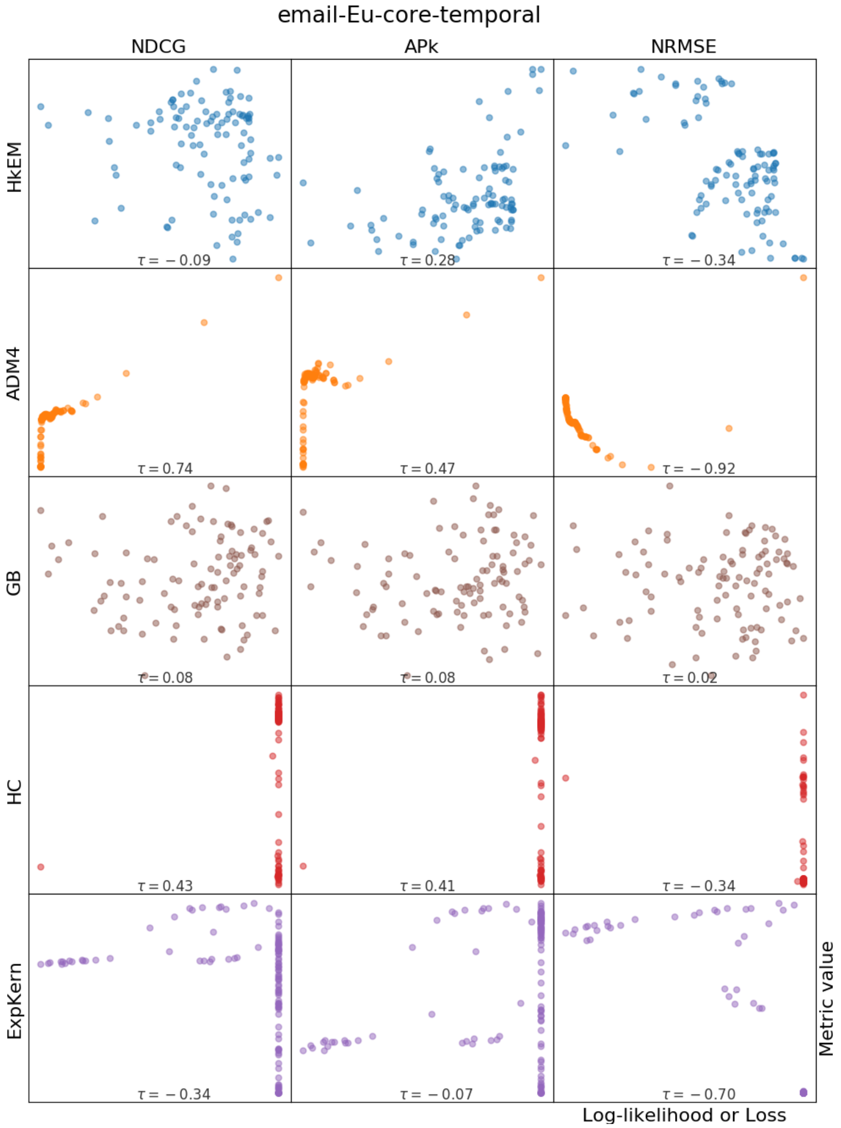
1.5 Additional plots for the correlation between log-likelihood and metric score

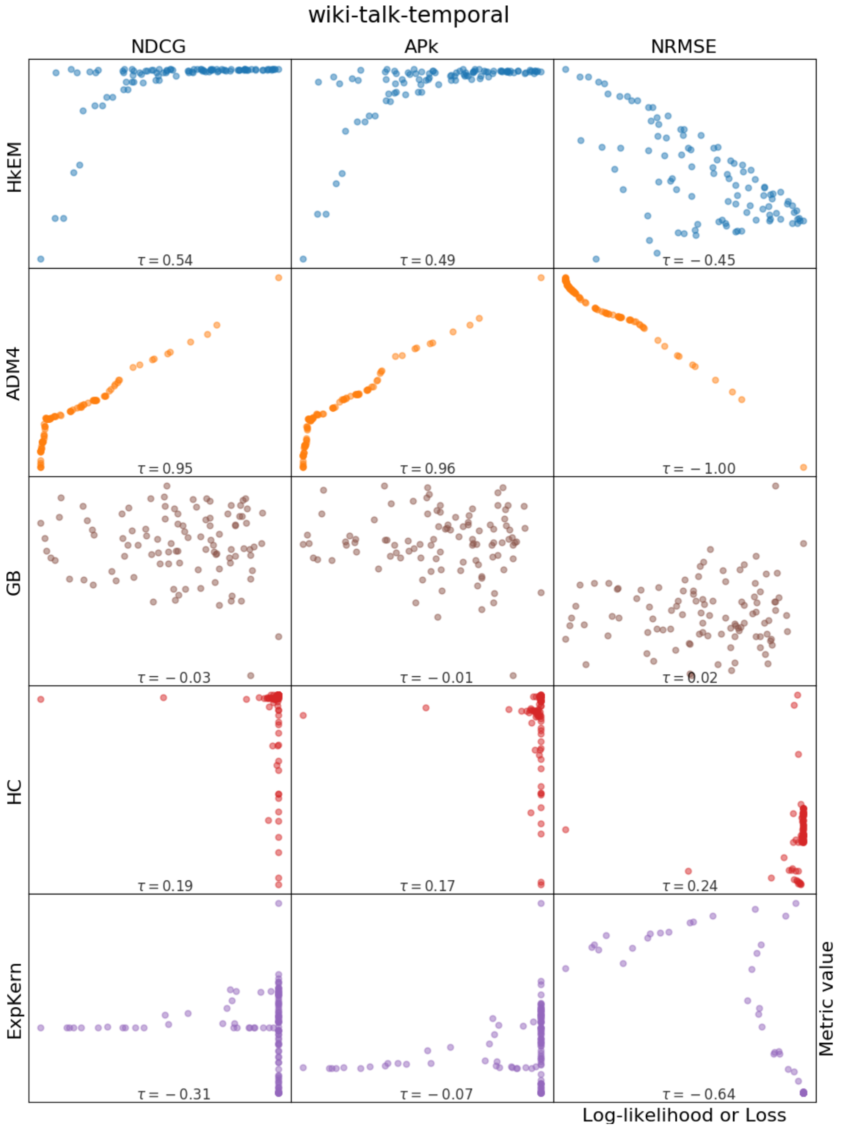

sx-mathoverflow
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
$\tau = 0.77$
$\tau = 0.76$
$\tau = -0.77$
$\tau = 0.91$
$\tau = 0.91$
$\tau = -0.96$
$\tau = 0.10$
$\tau = 0.10$
$\tau = -0.05$
$\tau = -0.30$
$\tau = -0.25$
$\tau = 0.06$
$\tau = -0.35$
$\tau = -0.27$
$\tau = -0.44$
Metric value
Log-likelihood or Loss


sx-askubuntu
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
τ = 0.62
τ = 0.64
τ = −0.74
τ = 0.94
τ = 0.94
τ = −0.89
τ = 0.01
τ = 0.03
τ = −0.15
τ = −0.35
τ = −0.15
τ = 0.41
τ = −0.46
τ = −0.35
τ = −0.66
Metric value
Log-likelihood or Loss


sx-superuser
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
$\tau = 0.69$
$\tau = 0.73$
$\tau = -0.67$
$\tau = 0.94$
$\tau = 0.94$
$\tau = -0.98$
$\tau = 0.03$
$\tau = 0.05$
$\tau = 0.05$
$\tau = -0.29$
$\tau = 0.04$
$\tau = 0.25$
$\tau = -0.29$
$\tau = -0.21$
$\tau = -0.64$
Metric value
Log-likelihood or Loss


CollegeMsg
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
τ = 0.51
τ = 0.30
τ = 0.20
τ = 0.98
τ = 0.96
τ = − 0.85
τ = − 0.04
τ = − 0.07
τ = 0.05
τ = 0.43
τ = 0.37
τ = − 0.51
τ = − 0.57
τ = − 0.25
τ = − 0.68
Metric value
Log-likelihood or Loss


email-Eu-core-temporal
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
τ = −0.09
τ = 0.28
τ = −0.34
τ = 0.74
τ = 0.47
τ = −0.92
τ = 0.08
τ = 0.08
τ = 0.02
τ = 0.43
τ = 0.41
τ = −0.34
τ = −0.34
τ = −0.07
τ = −0.70
Metric value
Log-likelihood or Loss


wiki-talk-temporal
NDCG
APk
NRMSE
HkEM
ADM4
GB
HC
ExpKern
τ = 0.54
τ = 0.49
τ = −0.45
τ = 0.95
τ = 0.96
τ = −1.00
τ = −0.03
τ = −0.01
τ = 0.02
τ = 0.19
τ = 0.17
τ = 0.24
τ = −0.31
τ = −0.07
τ = −0.64
Metric value
Log-likelihood or Loss

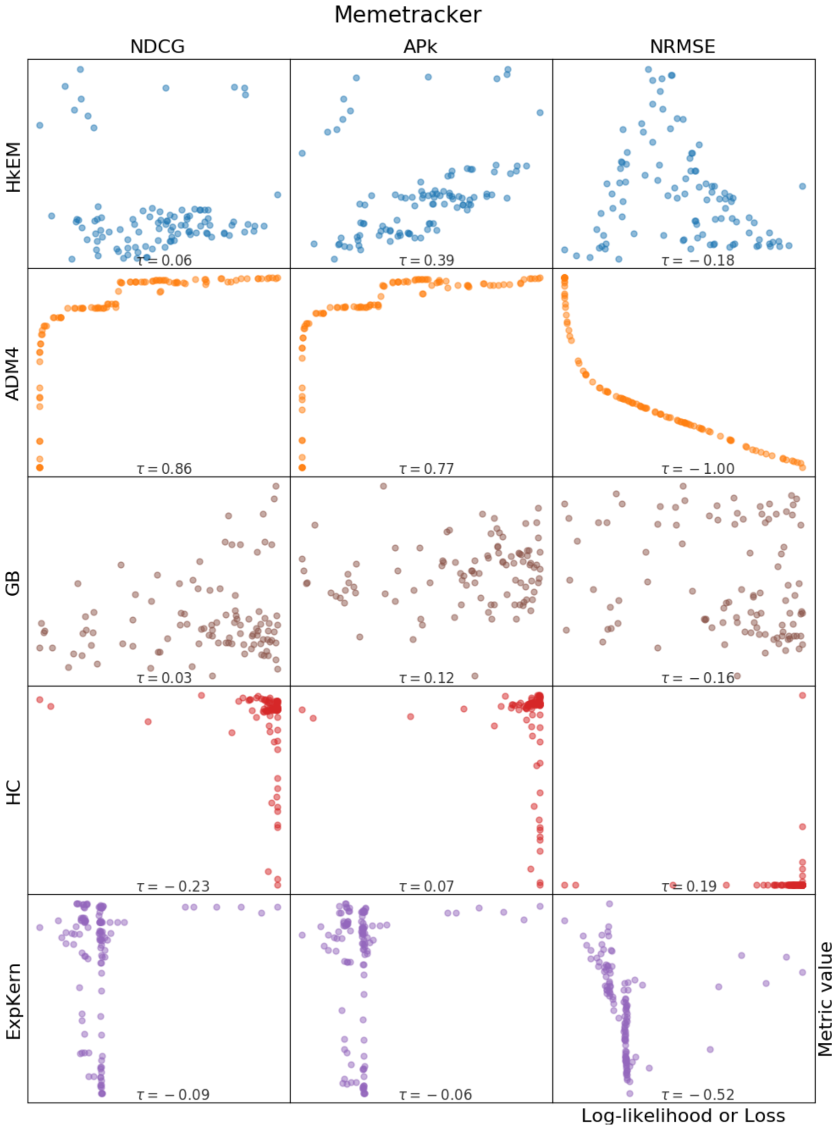


Fig. 3. Log-likelihood of the model (or loss function for HC) vs metric values when comparing the generated network with the ground truth for all datasets.

References

1. M. Achab, E. Bacry, S. Gaiffas, I. Mastromatteo, and J.-F. Muzy. Uncovering causality from multivariate hawkes integrated cumulants. In *ICML*, 2017.
2. F. Figueiredo, G. R. Borges, P. O. V. de Melo, and R. Assunção. Fast estimation of causal interactions using wold processes. In *NeuRIPS*, pages 2975–2986, 2018.
3. M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, pages 239–251, 1945.
4. H. Xu, M. Farajtabar, and H. Zha. Learning granger causality for hawkes processes. In *ICML*, 2016.
5. K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649, 2013.