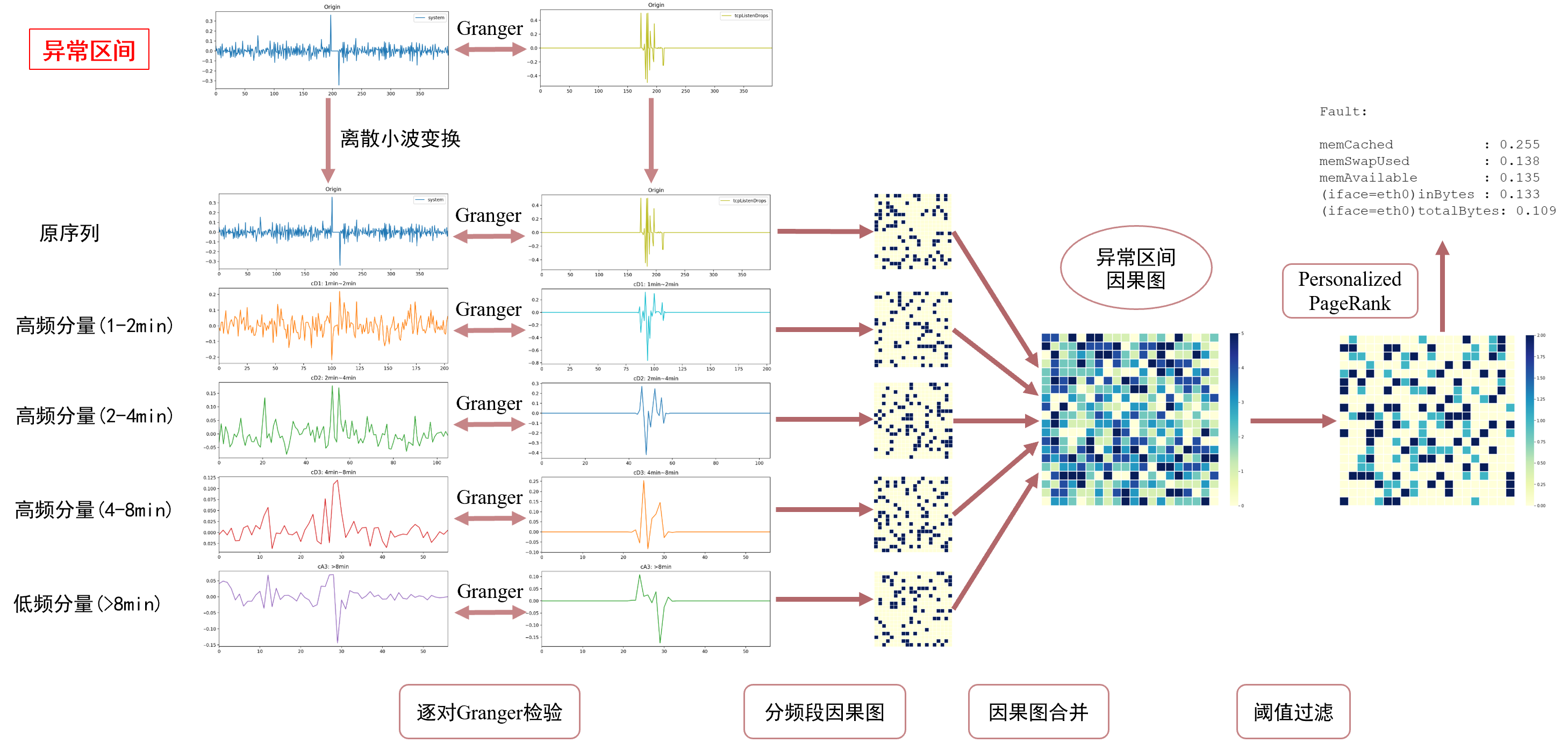
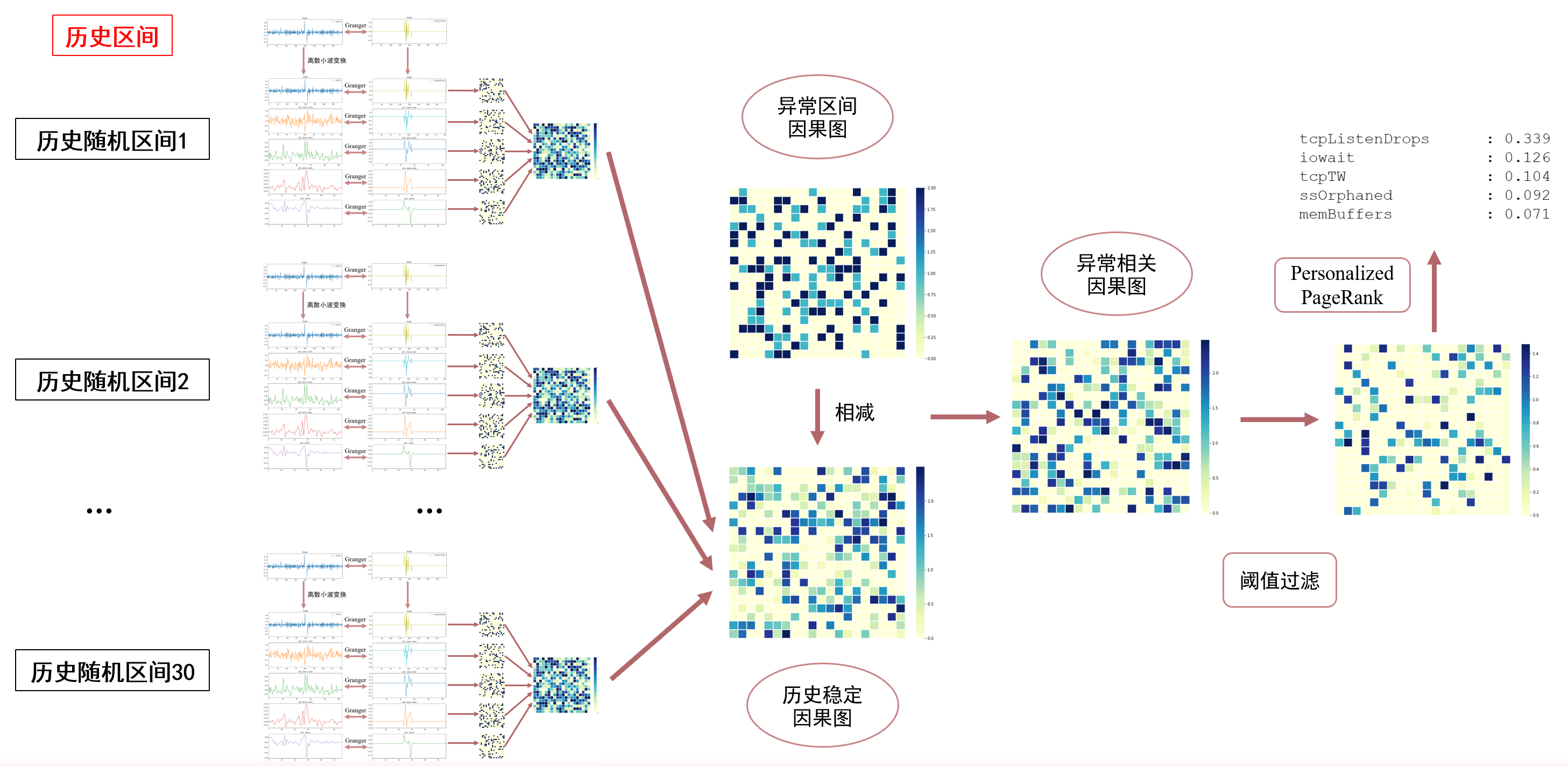
# 结合历史稳定关系的多尺度Granger分析

## 算法思路

监控指标序列在不同时频尺度上往往会展示出不同特性。对于许多序列而言，低频成分通常蕴含了该序列的固有特征，而高频成分则给出了序列的细节或差别。我们将异常时段的监控指标进行多尺度分解[1]，在各尺度上分别生成点对点的Granger因果关联关系图。我们将多尺度关联图合并为异常区间关联图，在图上通过Personalized PageRank[4]算法进行随机游走并得到平稳概率分布，根据概率值高低给出指标推荐。

考虑到稳定的因果关联关系对结果的影响，我们在历史时段中随机选多个历史区间（典型值为30个），在每个历史区间上都进行多尺度分解并生成多尺度关联关系图。这些关联关系图在多个历史区间上的平均构成了稳定关联关系图，它表征了不同指标之间内在固有的因果关联关系。考虑到这种固有的因果关联关系对PageRank算法的随机游走有干扰，我们在各尺度上提取稳定关联图和异常区间关联图的公共边，并将它从异常区间关联图中去除。通过合并多尺度的异常区间关联图，我们最终得到异常关联图，图上的边代表指标之间关于异常传播的关联关系。我们同样用Personalized PageRank算法得到平稳概率分布，并给出指标推荐。实验结果证明了去除稳定因果关联关系对算法准确度有大幅提升。





## 技术细节

### 监控指标多尺度分解

算法通过离散小波变换[1]将给定区间内的监控指标序列分解为多尺度时间序列。考虑异常传播的特性，将监控指标序列分为三个高频分量和一个低频分量：cD1（对应时间尺度为1~2min的高频成分），cD2（对应时间尺度为2~4min的高频成分），cD3（对应时间尺度为4~8min的高频成分）和cA3（对应时间尺度大于8min的低频成分）。高频分量包含序列细节信息，捕捉序列频繁的波动模式，时间尺度越小的高频分量越能捕捉振荡更剧烈的局部特征。低频部分是原序列去除了局部高频信息后的整体特征，表征了序列在较大时间尺度上的数据变化情况。经过监控指标多尺度分解，可以得到原序列、三个高频分量以及一个低频分量这五个频段的对应序列。这五个频段分别记为。

### 异常区间多尺度关联图生成

给定异常时间点，我们选取出该点前后各200个时间点的所有指标序列，通过前述的多尺度分解方法得到对应的多尺度序列集合。对于每个尺度，我们两两计算指标间的Granger因果性[2]： 。这样异常区间每个尺度上可以得到一个Granger因果关联图，表示在该频段对应的时间尺度内，指标之间的关联关系，记为。它表示了通过异常区间上的指标数据反映出的因果关联关系。

### 历史区间多尺度稳定关联图生成

同一主机在异常区间之前的时段称为历史时段。我们在历史时段中随机选取个历史区间，对于区间我们得到该区间上的指标序列以及对应的多尺度序列集合。和异常区间一样，我们在该区间上两两计算Granger因果性。对于所有的历史区间，我们对多尺度关联图做平均，得到历史区间的多尺度稳定关联图。它表示在正常的时段中，指标之间的稳定因果关联关系。越接近，表明在尺度上，指标和指标的因果关联关系出现得越频繁，表现得越稳定。实验中我们选取以达到算法稳定性和时间效率得平衡。

### 多尺度异常关联关系提取与合并

令，它表示排除了在尺度上指标之间的稳定关联关系后，与异常相关的关联关系的强弱程度。将所有尺度的异常关联关系相加合并，得到总体的异常关联图。设定阈值，保留权值大于的边，得到最终的异常关联图。

### 异常关联图上随机游走及排名

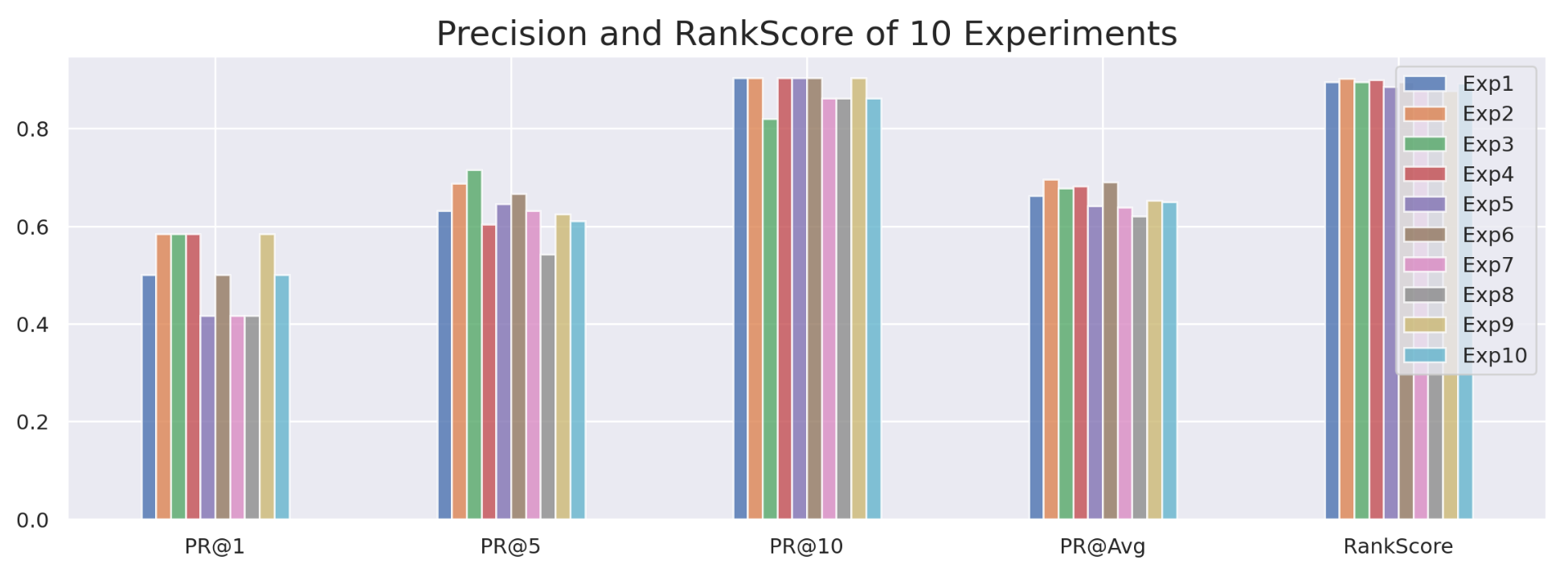
在异常关联图的反向图上，我们使用PageRank[3]算法为指标的重要性排名。PageRank算法基于图上的随机游走模型，通过表示一阶马尔可夫链的平稳分布得到图上各节点的平稳转移概率，并将这个概率作为节点的重要性。我们采用Personalized PageRank[4]算法，令概率分布向量(其中表示观测到异常的前端指标编号)，它表示了每次随机游走重新开始时的初始节点分布。通过概率分布向量，我们指定图上随机游走的指标入口，确保PageRank结果的有效性。我们按照各指标的平稳转移概率降序排名，并给出个指标及对应的概率值作为最终的推荐结果。

## 实验效果

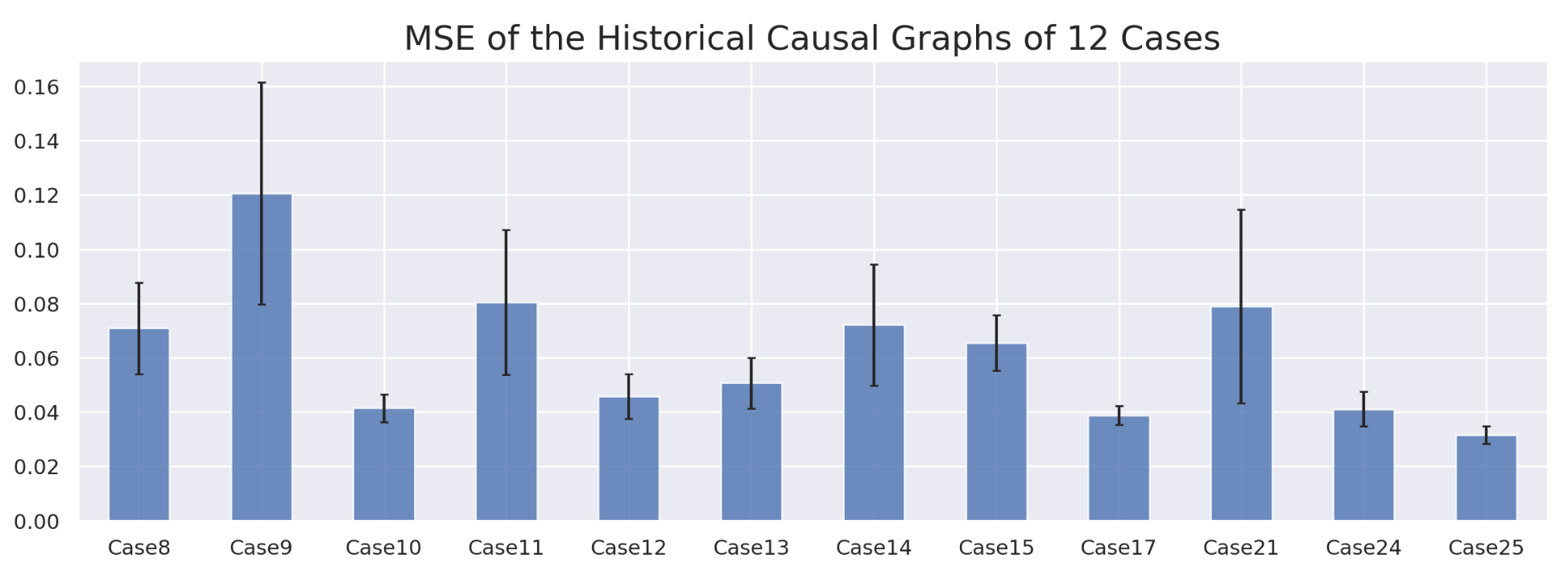
在12个case上，我们确定了准确的异常时间点、观测到的异常指标以及经运维人员确认的根因指标集。我们通过PR@k，PR@avg和RankScore性能指标衡量算法给出推荐的准确程度，并给出这些指标在所有case上的平均结果。由于算法选取历史区间具有随机性，我们重复10次实验并取结果的平均值。OnlyFault表示仅使用异常区间数据，不使用历史区间数据的算法结果，它衡量了算法在没有滚动保存历史指标序列的主机上的性能。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **PR@1** | **PR@5** | **PR@10** | **PR@avg** | **RankScore** | **Time(s)** |
| **OnlyFault** | 0.250 | 0.521 | 0.819 | 0.535 | 0.855 | 42.131 |
| **Ours** | 0.508 | 0.636 | 0.882 | 0.661 | 0.890 | 1037.413 |

算法在10次平均实验下的各指标变化幅度如下：



在10次平均实验下生成稳定因果关联图的MSE(Mean Square Error)分布如下：



## 参考链接

1. Sundararajan, D. (2016). *Discrete wavelet transform: a signal processing approach*. John Wiley & Sons.
2. Benhmad, F. (2012). Modeling nonlinear Granger causality between the oil price and US dollar: A wavelet based approach. *Economic modelling*, *29*(4), 1505-1514.
3. Gleich, D. F. (2015). PageRank beyond the Web. *siam REVIEW*, *57*(3), 321-363.
4. Bahmani, B., Chowdhury, A., & Goel, A. (2010). Fast incremental and personalized pagerank. *arXiv preprint arXiv:1006.2880*.