一、概念理解方面的问题

在文章《基于溯源图和注意力机制的 APT 攻击检测模型构建》中，对溯源图的处理及应用有如下问题：

1.文中使用的攻击事件序列，并没有体现实体间的因果关系

文中对溯源图的使用方法是，利用已知攻击节点（恶意进程、文件等实体），针对攻击节点，提取其在溯源图中的邻接图中的全部事件，构成攻击事件序列。如下图所示的溯源图中，实体 E 为标注的攻击节点，提取其邻接图，并按照时间序列构建攻击序列如下：

<C, create, E, T6>
<F, connect, E, T8>
<E, write, G, T9>

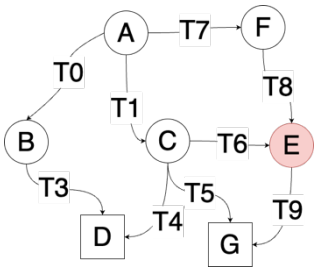


图 1

我认为该文章并没有用到图结构的因果关系，使用的是事件，仅能关注到一条边及其连接的两个实体节点。

2.按文章的处理思路，为何要还原溯源图这一步呢？

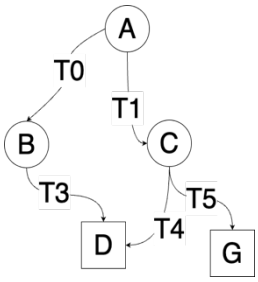
文中使用的是攻击序列，直接从日志中提取涉及攻击节点的事件然后按时间序列排好就行了，因为在其后续处理过程中完全没有考虑溯源图的存在，仅仅使用了攻击事件序列。例如假设图一对应的从日志提取出的事件内容如下：

<A, create, B, T0>;
<A, create, C, T1>;
<B, connect, C, T2>;
<B, write, D, T3>;
<C, write, D, T4>;
<C, write, G, T5>;
<C, create, E, T6>;
<A, create, F, T7>;
<F, connect, E, T8>;
<E, write, G, T9>;

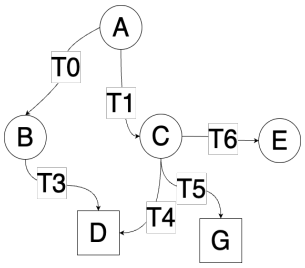
现在知道 E 被标注为攻击实体，直接从这个构成溯源图的事件序列中把包含实体 E 的子序列抽出来就构成所需的攻击序列了，完全不需要按原文所说还原溯源图后对溯源图进行压缩等操作。文中这样先还原成溯源图的目的是什么？

3.动态更新溯源图可行吗？

假设 T5 时刻，系统的溯源图为图 2（1）所示，T6 时刻系统溯源图为图 2（2）所示，



（1）



（2）

图 2

利用 GNN/GCN 获取图 2（1）所示的图特征向量x_5，图 2（2）所示的图特征向量x_6,以此类推，构成图特征向量序列$S = \{x_1, x_2 \ldots\ldots, x_5, x_6 \ldots\ldots, x_n\}$作为 transformer 的输入，对下一个时刻图结构进行预测，并对预测结果进行分类，预测攻击事件的发生。

这样的思路有可行性吗？

二、数据集方面的问题

前文提及的文章中使用的数据集来自《ATLAS: A Sequence-based Learning Approach for Attack Investigation》，其使用的是基于漏洞 CVE-2017-0199 自行模拟生成的系统日志、DNS 记录、浏览器日志。对于数据集有如下问题：

1. 如果日志中出现关键元素缺失，如何处理

例如，在浏览器日志中，某条日志缺少事件四元组<src, action, target, timestamp>中的 target，该条日志记载了 src 调用了其内部的某个方法，没有说明 target 是谁，对于此类日志应当如何处理，是直接舍去，还是保留？若需要保留，事件中的 action 在溯源图中表示为边，单这条边没有指向对象，应该如何处理？是否应该指向自身？