2016 年第三届中国可视化与可视分析大会

数据可视分析挑战赛

(ChinaVis Data Challenge 2016)

挑战 1

背景介绍

WeSuCom 是一家知名的通信服务公司，近年来一直致力于为企业和政府机构提供定制化的通信服务与应用解决方案。2015 年 7 月，WeSuCom 公司为某商业集团股份有限公司（BigBusiness）部署了一套网络化业务支撑平台，同时还部署了一套最新研发的网络监控系统，该监控系统在 BigBusiness 公司的骨干通信链路上对数据包信息进行抓取，特别是它可以记录数据包在链路层、网络层和应用层的相关信息。

通过一段时间试运行，WeSuCom 公司想了解新上线的业务支撑平台的运行状态，同时也想了解新研发的网络监控系统到底能帮助网络管理人员从监控日志中发现哪些有趣的模式，如何从应用层、网络层和链路层三个方面抓取的数据包信息中探索用户行为模式。但分析和理解大量包含丰富属性的网络监控日志是一个巨大挑战，因此，WeSuCom 公司提供了试运行期间两个月的网络监控日志，希望参赛者从网络管理人员的角度，采用可视分析技术对数据进行分析，帮助公司找到上述问题的答案。

数据说明

本次提供的网络监控日志数据，其时间跨度为 2 个月，共有 200 多万行记录。与只关注网络层活动的传统 tcpflow 数据不同的是，该监控日志是将网络数据包在应用层、网络层和链路层的相关信息抓取下来，更为丰富和全面，可以从多层次多角度反应数据在网络中的流动过程。

在 IP 网络中，应用层负责将需要传输数据拆分成一个或多个数据包，网络层负责在源地址和目的地址间建立逻辑连接，链路层负责为逻辑连接建立实际的传输通道，每次成功的网络连接都会完成若干数据包的传输。光纤是链路层最典型的传输介质，一根光纤通常集成了许多的通信线路，每个数据包都是经过光纤上不同的“虚拟管道”进行传输的。可以将光纤理解为许多电话线绑定在一起，每个数据包都通过某根电话线进行传输。以 155M 带宽的光纤为例，在去除网络管理部分占用的带宽之后，通常分成 3 个 45M 的通信线路，每个 45M 的通信线路又会按照传输的数据是网页、语音等不同分成不同的“虚拟管道”。

网络监控系统被部署在 BigBusiness 公司网络的骨干线路上，它可以监控所有通过骨干线路传输的数据包，通常跨子网的数据传输都要途径骨干线路。在生成日志时，网络监控系统会尽量将属于同一网络连接的数据包信息进行组装，并尝试解析在网络连接中传输的是什么类型的数据。因此，监控日志中每一行记录描述了一次网络连接数据传输在链路层、网络层和应用层的相关信息，具体字段如下：

- 应用层数据项包括：ID（记录序号）、IPSMALLTYPE（IP 业务类型，主要包括 66 和 67，分别代表了 TCP 和 UDP 协议）、FILELEN（该次网络连接所传输数据的总长度）、FILEAFFIX（此次网络连接可能传输的文件类型，若为 unk 表示未知文件类型）、ISCRACKED（监控系统判断数据包是否有损坏）共 5 个维度。
- 网络层数据项包括：STARTTIME（开始时间），SRCIP（源 IP），DSTIP（目的 IP），SCRPORT（源端口），DSTPORT（目的端口）共 5 个维度。
- 链路层数据项包括 VPI1、VCI1、ATMAAL1TYPE 共 3 个维度，VPI 和 VCI 是链路层“虚拟管道”的标识，ATMAAL1TYPE 是这个虚拟线路中传输数据的属性。

以用户 A 发送一封邮件给用户 B 为例，该邮件首先在应用层被拆分成多个数据包，然后在网络层建立逻辑网络连接，最后这些数据包在光纤链路层介质中经过“虚拟管道”进行实际传输。上述三个层次的具体过程为：

（1）该份邮件在应用层会被分解成多个数据包进行传输，多个数据包可组成一个网络连接。WeSuCom 公司部署的网络监控设备会尽量将属于同一个连接的数据包信息组装起来，但是也存在组装失败的情况，因此，日志数据中可能存在一些表示离散数据包的记录，这些未能有效组装的离散数据包可能是由于数据丢失或者初始运行的错误产生的。ID 记录序号是递增的，但序号不一定连续，中间可能会由于组装连接失败或者丢失数据而不连续；FILEAFFIX 是此次连接可能传输的文件类型（目前大部分是未知 unk，也有一些 exe、bmp、jpg 等类型的数据）；ISCRACKED 标识了此次连接中是否有数据包被损坏，该标识可以反应出此次通信的质量。

（2）按照应用层协议规则处理之后，该邮件的数据包进入了网络层，网络层是以 IP 地址标识网络实体的位置，并以端口来区分不同的应用，即说明这封邮件的数据包要从网络中哪儿发向哪儿，记录的信息包括：源 IP（SRCIP）、源端口（SCRPORT），目的 IP（DSTIP）和目的端口（DSTPORT），以及时间戳 STARTTIME。

（3）一根链路层的光纤会包含多路“虚拟管道”，数据包走管道实现链路层传输，VPI1、VCI1 是“虚拟管道”的标识，ATMAAL1TYPE 给出了这个虚拟管道中传输数据的属性。通常用户 A 发给用户 B 的这封邮件的所有数据包都会走同一个管道。虚拟管道是可以动态配置的，例如：一路语音开始时占用了 64k 的虚拟管道，其功能类似于一条专门的电话线。在下一时刻，该路虚拟管道可能和另一个虚拟管道合并成 128k 进行传输，也可以多个合并来传输网页数据。

另外，由于设备原因或某些不可抗拒的原因，某些时段获取的数据记录可能是空的，即日志数据缺少某些时间的记录，这是进行流量分析时要考虑的因素。数据中也可能存在某些字段值采集异常的情况，常见的一种异常是某一行的记录中由于某个字段的数据没有来得及写入而造成该行部分内容的错位，这类数据可认为是噪声，分析时需进行剔除。

题目说明

（1）找出 BigBusiness 公司内部网络中的客户端与服务器，并给出 BigBusiness 公司的网络体系结构拓扑图；对 BigBusiness 内部网络中的服务器进行分类，分类标准不限，比如:按照节点类型、按时间特点、按行为特点、按流量特点等等。

（2）找出可能存在的异常通信模式（异常事件），异常标准不限，比如：网络中的周期行为模式和活跃时段变化、某个子网访问或被访问量变化、子网连接模式变化、连接可能传输的文件类型（exe、bmp、jpg 等）的变化等等，建议给出至少 5 种异常通信模式。

（3）找出经同一虚拟管道（VPI、VCI）进行通信的源 IP 和目的 IP 的分布规律和连接模式；找出经同一虚拟管道（VPI、VCI）承载的应用分布和变化规律；说明每一个（VPI、VCI）虚拟管道的传输数据量在不同时段的变化情况；建议至少给出 5 种虚拟管道通信模式。

（4）WeSuCom 公司要求参赛队设计的可视分析方案能展示数据从源到目的经历应用层、网络层和链路层的全过程，实现对数据连接的多层次可视分析，探索其可能反应的用户行为模式。请参赛队在本题中具体说明其方案是如何满足该需求的。