1. 各位老师你们好，我是兰雨晴教授的学生王文博，我的毕设题目是恶意域名检测技术研究
2. 以下是第一部分的内容：研究背景和问题的提出
3. 根据2016年的中国互联网网络安全报告可以看出，虽然相关安全问题有所遏制，但是网络中恶意域名数量依然庞大
4. 在网络安全防护意识提升的同时，黑客的攻击手段也不断进化，例如这两年活跃的WannaCry、Mirai等，都造成了巨大的损失，其他的恶意域名还包含有DNS放大攻击相关域名、随机子域名攻击、色情域名等等。恶意域名危害不断增大
5. 有幸在中心的学习中接触到来自中国电信部门的被动DNS数据，论文中对于恶意域名的研究也全部基于这些数据。
6. 当前恶意域名的数量十分庞大、危害性增大，而对与被动DNS的分析还远远不足，如何利用好这些数据就成了我的研究重点
7. 接下来我将对国内外研究现状进行简述
8. 首先，在针对恶意域名的手段上主要有这三种方式，逆向工程、信誉系统和机器学习，他们都有这各自的缺点
9. 国外的研究人员最主要介绍的是Antonakakis，最为经典的信誉系统Notos和检测系统pleiades都是他们的论文成果；当前最为活跃的安全公司为Nominum
10. 国内也有很多研究机构和商业机构从事这方面研究，论文中使用的DGA黑名单均来自于360网络安全实验室
11. 接下来进行研究内容的介绍
12. 首先是特征提取的介绍，域名的特征我将其分为字符特征和访问特征，字符特征所针对是不包含顶级域名的二级域名部分，即图中的“example”
13. 具体来看，N-gram特征使用白名单计算分别计算n=1、2、3、4时的字符串概率，
14. 处理时将“example”的前后分别添加起始符和终止符，然后进行划分和计算。之后使用待检测的域名列表以同样方式处理域名，分别计算这4组概率，统计其均值、中位数、标准差，共12个N-gram特征
15. 图为黑白名单中域名长度和熵值的关系，按照这个公式计算熵值，除以长度之后得到字符平均熵，共2个特征
16. 其他的统计特征包含长度、字母和数字长度、顶级域名共4个特征
17. 域名访问特征分为访问统计特征和域名词向量特征
18. 访问特征如表所示，按照不同的查询类型和域名服务器的响应类型进行统计
19. 由此计算得到一些衍生特征
20. 随后是对这些特征的分析方案，针对衍生变量，通过计算之间的相关系数，观察是否和原变量具有线性相关性
21. IV值是对特征预测能力强弱的一个判定标准，IV值的计算是由WOE得到，Woe的计算如此表
22. 对每一个woe加权之后累加即得到IV值
23. 此表为IV值与预测能力之间的关系
24. 下面是恶意域名相关流量提取的算法流程，根据Nominum的报告，当前网络中占比最大的三类恶意流量为放大攻击流量、随机子域名攻击流量、DGA流量，总体占据99%以上，因此本论文以信誉系统为蓝本针对这三类域名提出了流量提取的算法方案，保证召回率的前提下，将被动DNS数据量降下来
25. 数据预处理包含了数据清洗和数据过滤，数据清洗包含去除这两类情况、数据过滤包含这三类情况
26. 针对DNS反射放大攻击的相关域名提取中，主要使用了TXT查询比例和ANY查询比例，构建公式
27. 针对随机子域名相关域名的提取主要使用二级域名下子域名计数和不存在域名比例构建公式
28. 针对DGA域名的提取使用长度、熵值、N-gram特征，使用随机森林分类
29. 基于词向量空间的色情域名检测算法，创新的将自然语言处理的词向量算法用于色情域名检测当中，和以往针对内容和用户日志的算法完全不同。具体流程如图，将流量处理为IP-domain对构建语料，得到域名访问序列，使用Word2Vec框架无监督训练得到域名词向量，以这些词向量作为特征进行分类的实验
30. 语料的组织如图所示，将同一IP的域名按照访问序列组织起来
31. 下面是多维度特征的DGA域名检测算法，使用了上面提到的字符特征和访问特征，
32. 域名特征组织如图所示
33. 接下来是实验结果和分析
34. 山西省数据为2016年10月15日 23个小时的数据， DNS记录的总量接近20亿条，每小时不重复的二级域名数量在10万到20万之间，共标注DDOS相关恶意域名101个，DGA相关域名322个。广东省的数据为2017年4月14日 9点到16点共计9个小时的数据，DNS记录总量达11亿余条，每小时不重复的二级域名数量达到60万左右，共标注DDOS相关恶意域名163个，DGA相关域名265个。
35. 图为α、β下召回率的影响，得到α=0.1，β=0.05，右边图为θ和域名数量关系，当θ取值大于0.2时，召回率达到100%。θ最终取值为0.3
36. 这是山西的结果。(a)图表示提取的随机子域名攻击涉及域名和DNS放大攻击涉及域名数量所占比例，(b)图表示提取的随机子域名攻击涉及域名和DNS放大攻击涉及流量所占比例，(c)图表示提取的DGA域名数量所占比例，(d)图表示提取的DGA域名流量所占比例。
37. 这是广东的结果。其中随机子域名和DNS放大攻击涉及域名提取的召回率达到100%，DGA域名的召回率达到92%。
38. 接下来是色情域名的样本分析，使用下采样的方法，对负样本进行了抽样，保留所有正样本。
39. 色情域名检测结果，使用27日作为训练、分别使用28、29日验证
40. 下面是DGA域名检测样本分析，同样对负样本进行了抽样
41. 接下来是特征分析结果，对域名词向量特征在DGA域名检测上的AUC和KS计算，auc为0.915，ks为0，649
42. 这是访问记录特征衍生特征和原始特征的相关系数，其中部分特征存在较强的现行相关性
43. 对这些访问特征计算IV值，观察预测能力得到51个预测能力较好的特征
44. 这是字符特征的IV值计算，当前word2vec特征的200维特征全部保留使用；164个访问统计相关的特征剔除预测能力差的以及冗余的特征后保留使用其中78维特征；18维的域名字符特征剔除其中预测能力较差的数字数量特真正保留剩余17个特征，共计295维特征。
45. 这是DGA域名样本外的验证结果和时间外验证结果
46. 最后是硕士期间取得成果
47. 感谢给位老师