数据可视化 分析说明文档

- 第六小组

小组成员： 徐悦皓、陈雨豪、萧升晓

题目来源： ChinaVis 2016 – Challenge 2

**题目要求**

Hacking Team是一家来自意大利米兰的信息技术公司，该公司向政府部门及执法机构提供信息系统入侵与监视服务，它帮助客户截获Internet用户间的通信、解密文件、监听Skype等网络通话、甚至还可以远程开启麦克风和摄像头。2015年7月5日， Hacking Team公司的官方Twitter遭不明人士入侵，被入侵后的首条通告写道：“反正我们也没什么东西好藏，那就把我们的电子邮件、文件、源代码都发布出来…”，随后公司大量内部数据被公开发布到网络上。这次特殊的数据泄露事件引起了社会各界的广泛关注，其中一项热门议题是如何解密Hacking Team公司的组织结构和发展历程。遭泄露的Hacking Team公司内部邮件数据是了解该公司的重要数据源，我们对邮件数据进行了初步的格式化处理，但分析和理解这批邮件数据仍然是一项非常艰巨的任务。因此，我们将格式化后的邮件数据提供出来，希望参赛者以数据分析师的身份，采用可视分析技术来分析邮件数据，帮助我们了解Hacking Team公司发展历程及各阶段业务特点，找出该公司内部的重要人物并推理其担任的角色与工作职责。

**题目说明**

(1) 从邮件数据中找出Hacking Team公司**内部员工列表**，并尝试**对员工进行分类**，分类标准不限，可以同时综合考虑多种分类方式，比如：按员工在公司的重要程度分，按员工在公司的角色分，按员工在公司的工作职责分，或按员工的行为特点分。

(2) **对邮件进行分类**，分类标准不限，比如：内部工作相关邮件、垃圾邮件、群发邮件、告警邮件、会议通知、非公司内部邮件等等，可以同时结合多种分类方式，比如：先按内部和非内部邮件分，然后再细分内部邮件。

(3) 根据邮件数据**总结Hacking Team公司经历了几个发展阶段**，每个阶段的**主要业务**和新增业务是什么，每个阶段的邮件数据中有哪些**热门话题**。

**零、 平台简单展示**

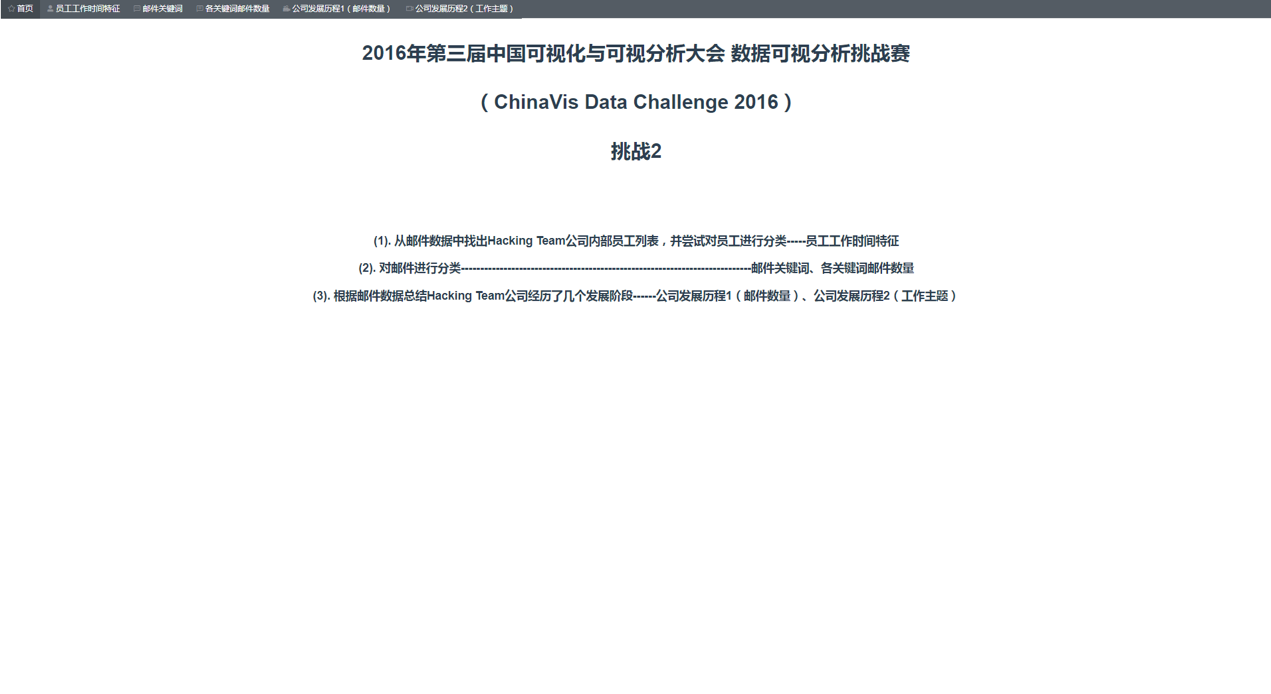


图0-1

图0-1 为本作品所假设平台的主页，其上方导航栏可跳转到各个“展示分析结果的可视化作品”所在的页面。

若要在本地打开该页面，可在解压src.zip压缩包后，打开”./src/display/index.html”文件即可。

也可选择直接访问网页进行查看：

”<http://tinghaode.ren/something/ChinaVis16-2/dist/index.html>”

本作品主要采取web形式展示，使用js及d3库，基于vue框架进行前端设计，源代码在”./src/data\_visualization”目录下。

在数据处理部分我们使用python语言，源代码在”./src/data\_processing”目录下。

**一、 分析人员**

（1） 筛选内部员工

筛选内部员工时，我们小组的判断标准为“使用hackingteam域名发送接收一定数量邮件的人为员工”，同时筛去了一些非人名的测试账号及相关邮件数量小于50的员工账号。

筛选后的“员工及其相关邮件数”列表如下：

|  |  |
| --- | --- |
| 员工名称 | 相关邮件数量 |
| marco bettini | 80638 |
| giancarlo russo | 58463 |
| massimiliano luppi | 46721 |
| antonella capaldo | 36454 |
| david vincenzetti | 29702 |
| lucia rana | 29170 |
| simonetta gallucci | 27246 |
| mostapha maana | 25835 |
| alessio scarafile | 23208 |
| fabio busatto | 22944 |
| marco valleri | 19617 |
| daniel maglietta | 19474 |
| alex velasco | 17959 |
| marco catino | 12948 |
| mostapha maanna | 12590 |
| emad shehata | 11348 |
| daniele milan | 8655 |
| serge wood | 8544 |
| bruno muschitiello | 7740 |
| massimo chiodini | 7335 |
| antonio mazzeo | 7237 |
| fulvio degiovanni | 7091 |
| sergio solis | 5844 |
| mauro romeo | 5815 |
| alberto pelliccione | 5652 |
| luca filippi | 5097 |
| ivan speziale | 4414 |
| guido landi | 3979 |
| walterandrea furlan | 3877 |
| eros marcon | 3625 |
| roberto banfi | 3610 |
| alberto ornaghi | 3166 |
| diego giubertoni | 2697 |
| daniele molteni | 1803 |
| max luppi | 1516 |
| cristian vardaro | 1248 |
| eric rabe | 1235 |
| philippe vinci | 1198 |
| ivan roattino | 1118 |
| giovanni cino | 1096 |
| matteo oliva | 981 |
| alessandra mino | 896 |
| alessandro lomonaco | 789 |
| danilo cordoni | 771 |
| stefania iannelli | 744 |
| christain pozzi | 706 |
| luca guerra | 686 |
| salvatore rumore | 667 |
| emanuele placidi | 483 |
| gianluca piani | 476 |
| enrico luzzani | 473 |
| alfredo pesoli | 471 |
| eduardo pardo | 424 |
| roby banfi | 422 |
| daniel martinez | 376 |
| andrea cariola | 375 |
| valeriano bedeschi | 355 |
| fabrizio cornelli | 330 |
| thomas valentini | 322 |
| marco fontana | 319 |
| gabriele parravicini | 319 |
| constantino imbrauglio | 307 |
| federico guerrini | 245 |
| massimiliano oldani | 211 |
| debora leanza | 209 |
| lorenzo invernizzi | 183 |
| claudio agosti | 176 |
| alessandro scarafile | 168 |
| eva michalikova | 166 |
| davide romualdi | 139 |
| sara galvagna | 115 |
| eugene ho | 97 |
| stefano bagnasco | 88 |
| aldo scaccabarozzi | 87 |
| sergio rodriguez | 63 |

（2） 对员工进行分类

对员工进行分类时，我们小组采取的分类依据为“一天当中的工作时间段”，具体通过每位员工发送邮件的时间段进行统计。

【原数据为东八区时间，数据处理时已将其转换回意大利所在的东一区时间】

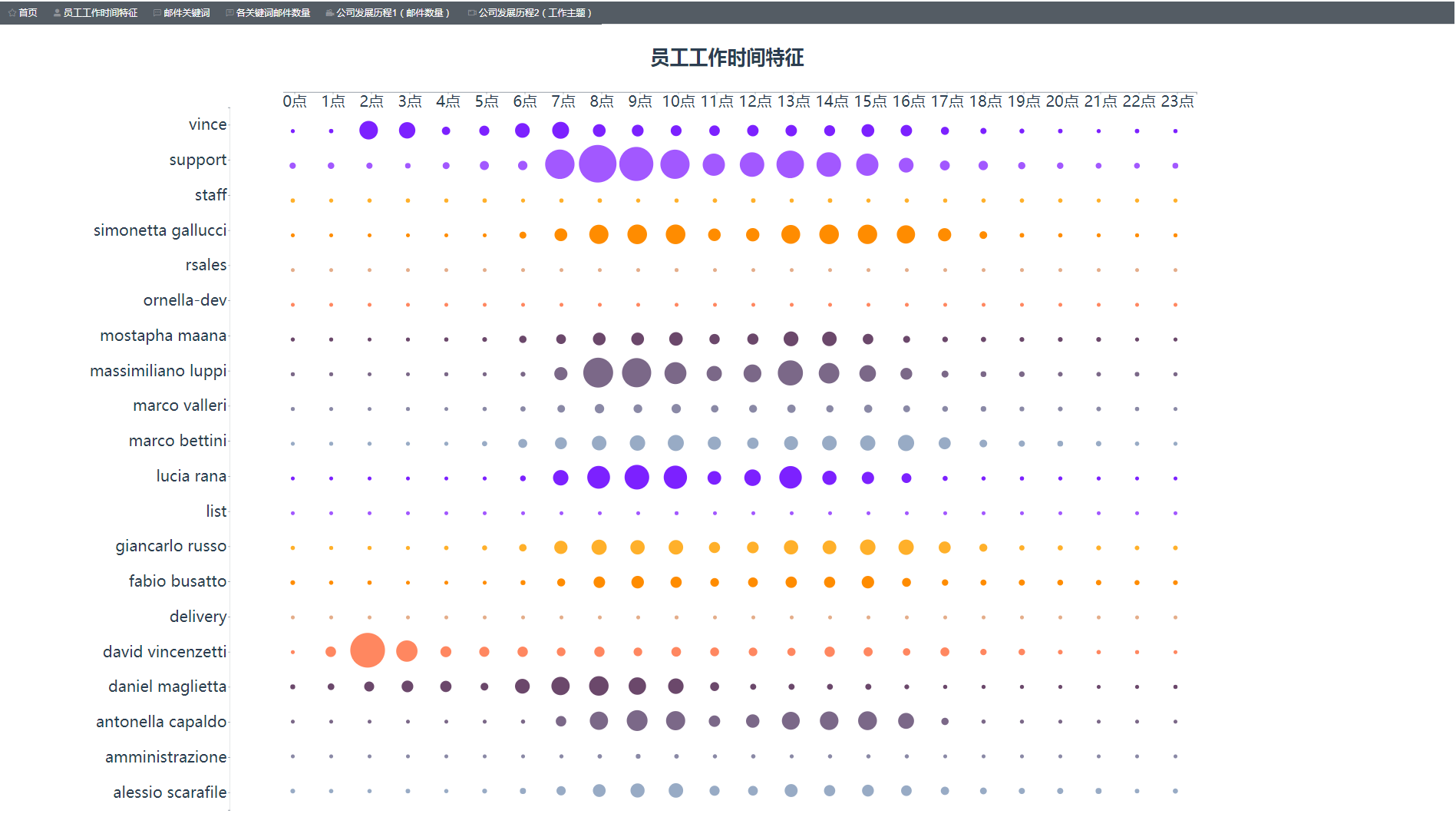


图1-1

图1-1 为员工工作时间段统计表。其中横轴是时间段（0-23小时），纵轴是前20个发送邮件数量最多的内部员工姓名。圆点越大，表示在该时间段内发送的邮件数量越多。

从 图中 可以看出大部分人发送邮件的时间集中在上午和下午，基本上是正常的工作时间。然而，David Vincenzetti在凌晨2点到3点发送的邮件最多，在大家都工作的时候，却不怎么发邮件，说明他的工作模式和别人有很大差异，我们推测他应该是 CEO之类的角色，负责制定计划、分配任务，但是不需要考虑工作细节。

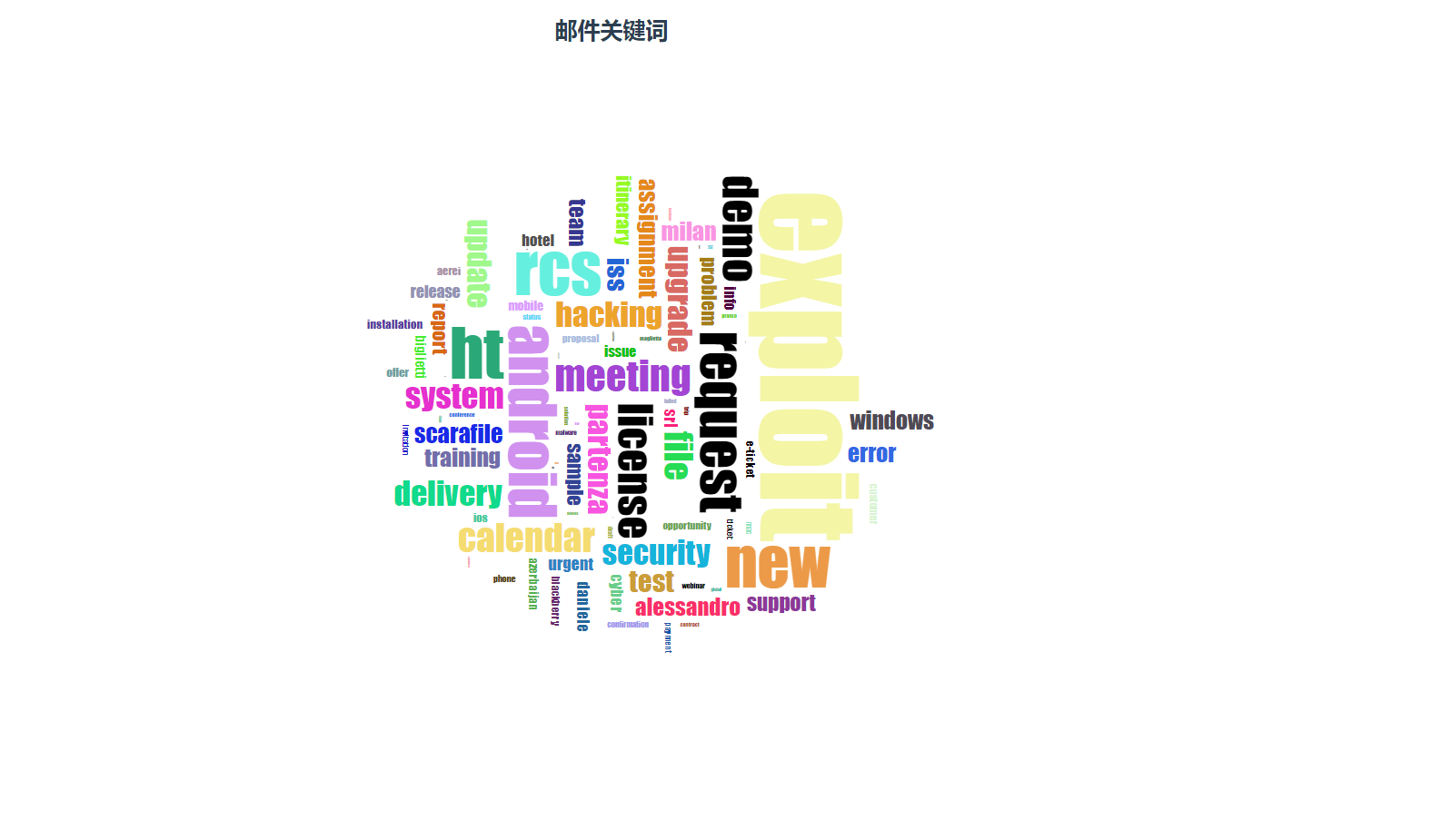
**二、 分析邮件内容**

在邮件内容的数据处理上，我们使用了包括 词形还原、统一大小写、停用词过滤等方法 对文本进行预处理。之后准备使用词袋模型进行特征处理并使用LDA模型进行主题分类，但出于多方面原因，最终只按照词频筛选出了最为重要的一部分关键词。

之后我们分两步对文本内容进行了可视化分析与处理。

（1） 使用 关键词词云 宏观感知

首先我们使用词云（如图2-1所示）了解整个关键词集合的大致特征，其中关键词越大，表明相关邮件的数量越多。

图2-1

（2） 使用 关键词分类sunburst图 具体描述

之后，我们对关键词做出了分类，将其分为“业务”与“广告”两类，并使用sunburst图对分类结果进行展示。当鼠标放至某一类时，该分类及其父分类将会亮起，上方标题栏也会显示出主题名称及相关邮件数量。

下面是其中几类的简单展示。

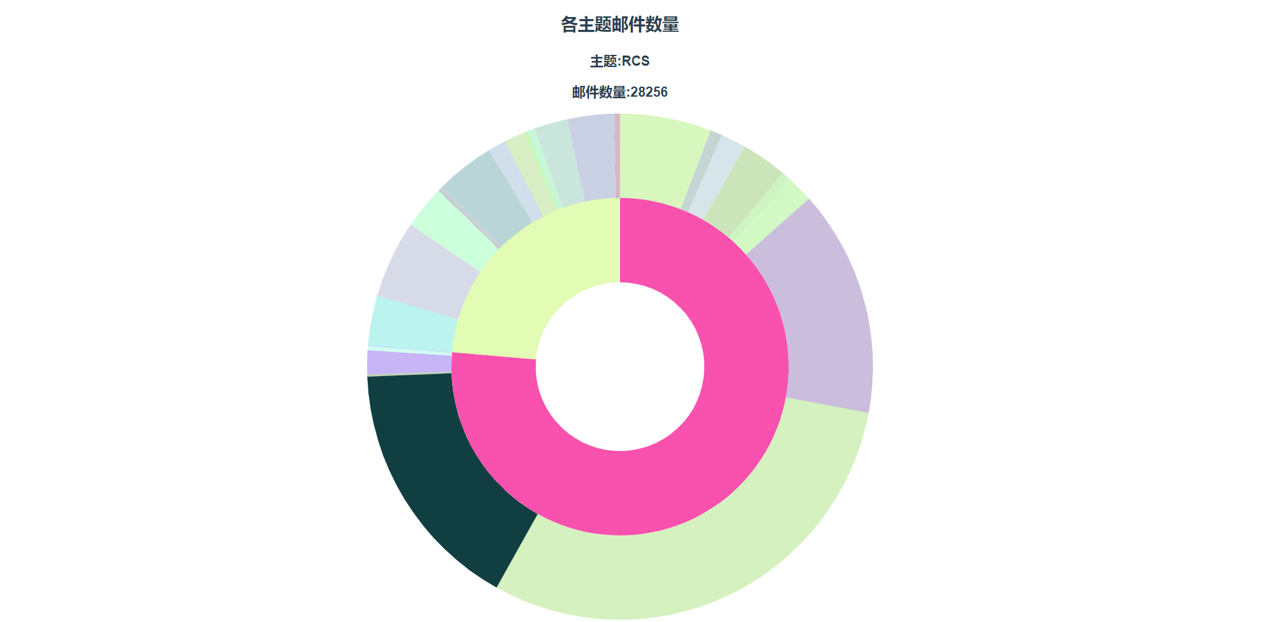


图2-2

从sunburst图中，我们可以看出：Exploit、RCS、Android、Windows等关键词频繁出现在邮件中，说明Exploit漏洞利用问题以及一些黑客工具、黑客消息是公司比较关注的核心问题，同时经查阅资料发现RCS为hackingteam公司最知名的产品，故在邮件中频繁出现也合乎逻辑；另外，Android可能是该公司的主要业务。

为方便下一题与业务相关的展示，我们又对业务进行了具体的细分，分为了“操作系统”及“攻击”两类。

每一类的具体内容及其中文含义如图2-3所示：



图2-3

**三、 分析公司历程**

在对公司历程的分析中，我们分别从“每年度邮件数量”及“每年度与关键词相关邮件数量”两个方面作出分析，分别达到“切分公司发展阶段”及“展示各阶段主要业务及热门话题”的目的。

（1） 每年度邮件数量

我们使用折线图展示了公司每年度的邮件数量，当鼠标放在折线图中的每个点上时，我们能看到该年度具体的邮件数量。

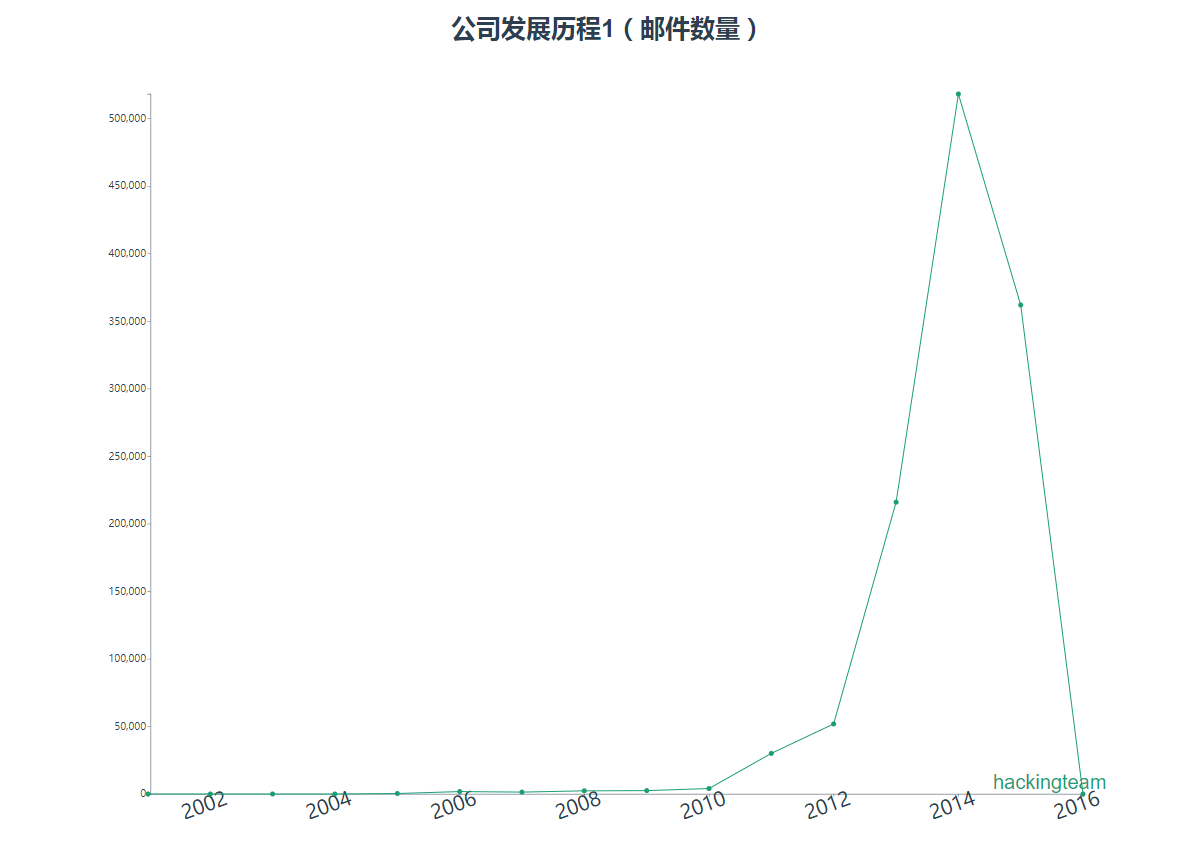


图3-1

图3-1中，横轴表示年份，竖轴邮件数量。

从折线图中我们可以清楚地看到：

在2001年到2005年，公司邮件数量非常稀少，此时应当是公司的起步阶段；

在2006年到2010年，公司邮件数量开始有了一定数量的爬升；

在2011年到2012年，公司有了稳定且较大幅度的增长，此时正是公司的发展期；

而2013年到2015年，整个公司突飞猛进，邮件数量达到了一个恐怖的量级，可见此时正是最火热的时期，但在2015年邮件数量下降，根据题目描述，这应当是数据泄露事件所导致的。

（2） 每年度 与关键词相关邮件 数量

我们使用相邻矩阵图展示每年度与关键词相关邮件数量，横轴代表年份，色块的颜色越鲜艳（红色）表明该关键词出现的频率越高。

同时我们在左边用饼状图标明了关键词的两个父类：“操作系统”/ “攻击”。当鼠标置于饼状图上时，同一分类的关键词将会亮起。

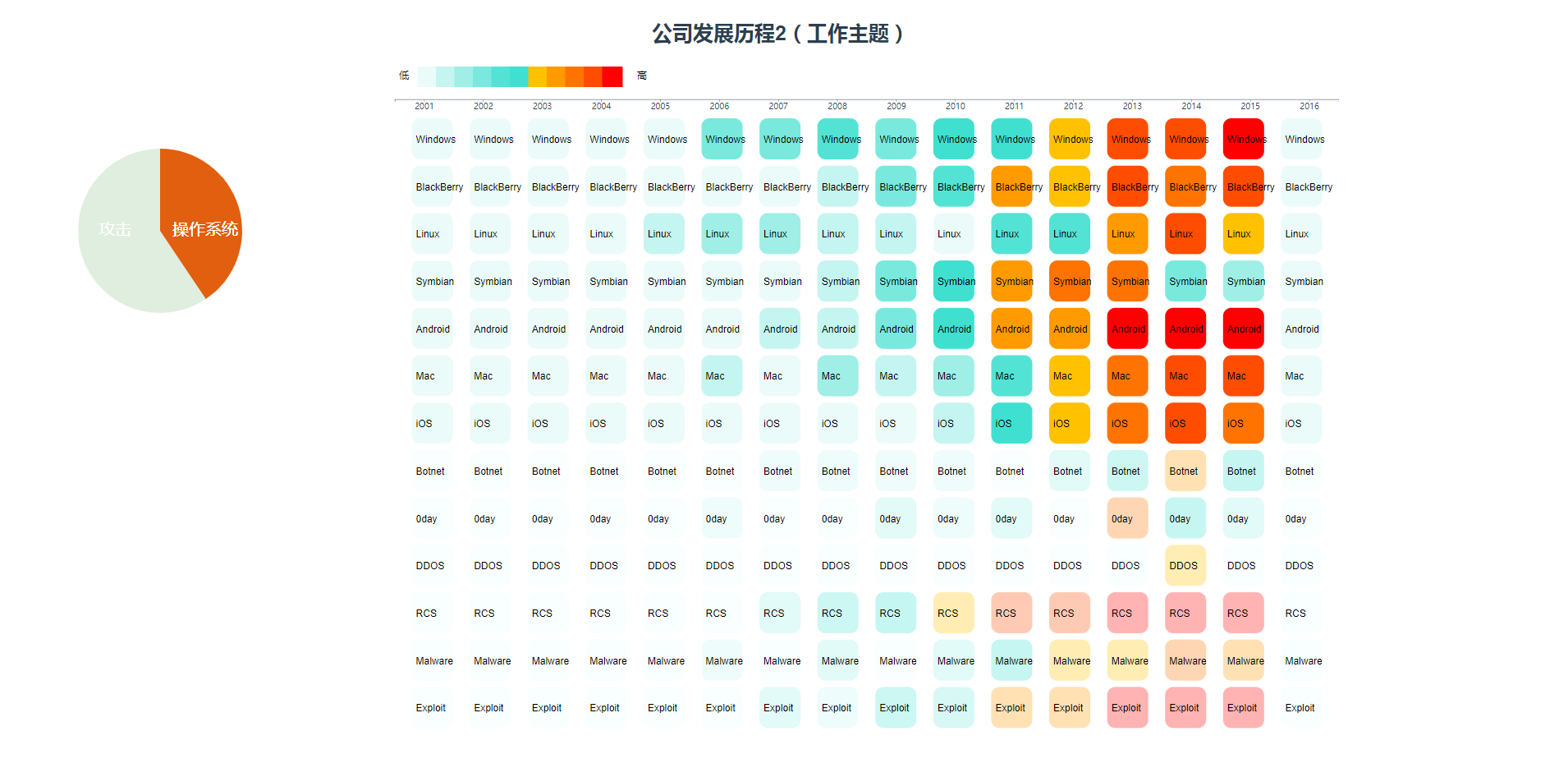


图3-2

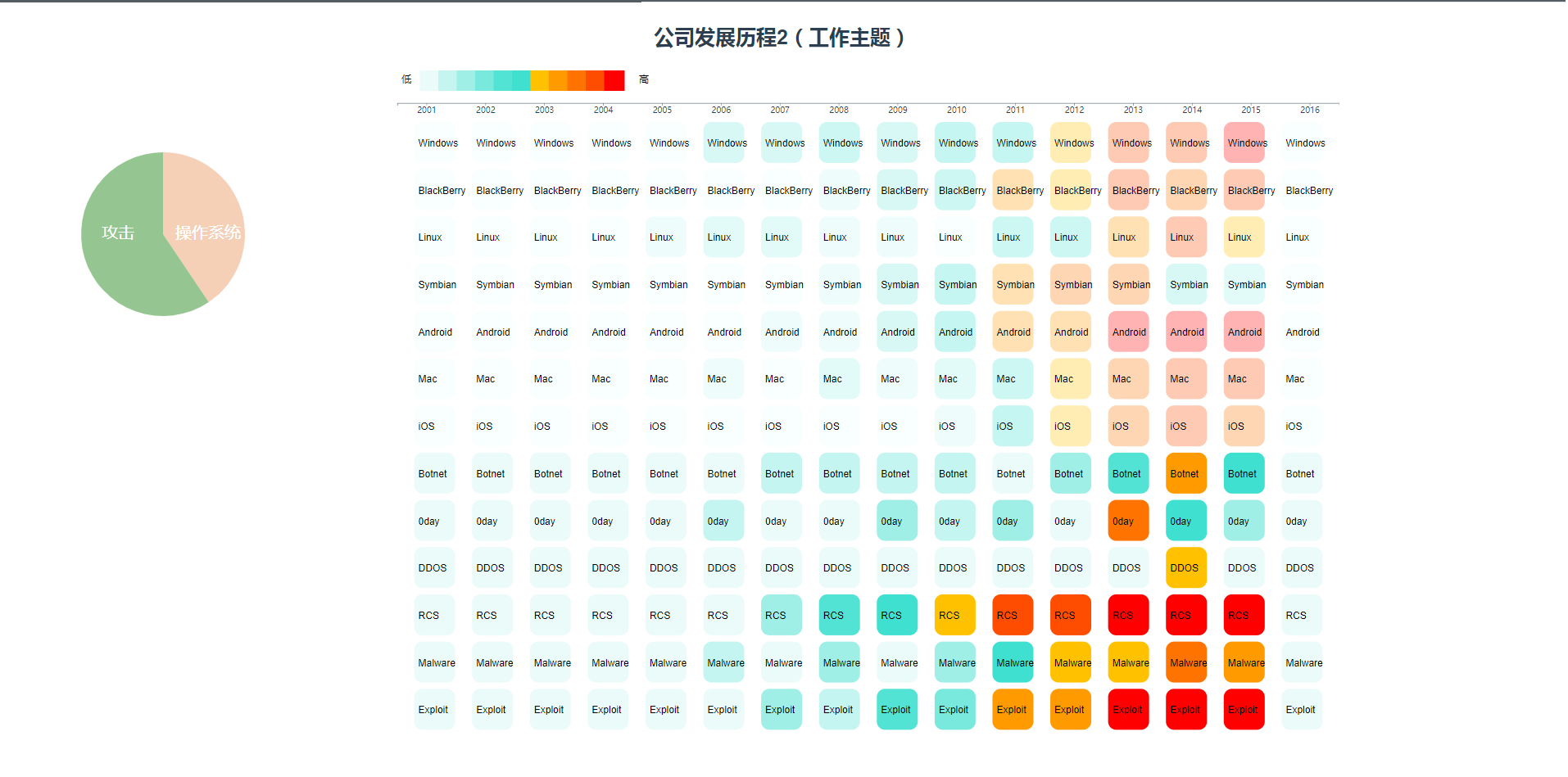


图3-3



图3-4

图3-2到3-4更进一步确认了我们之前的观点：

2001-2005年的色块普遍都很淡，而且从图3-1也可以看出2001-2005年期间Hacking Team公司往来邮件数量非常少，更表明了该公司正处于筹划期；以此类推，2006-2010年处于发展期；而 2010 年之后，Hacking Team公司进入了一个快速发展的时期，业务往来开始增加；在 2014年该公司的邮件数量最多，而且有多个颜色鲜明的色块，表明该公司处于一个顶峰时期；而到了2015年，该公司由于受到信息泄露事件影响，热度开始下降。

从exploit、cyber、system等红色块可以看出，网络安全监管、信息系统入侵应该是他们的核心业务。

在2001年到2005年，几乎没有主要业务，在Linux稍有涉足；

在2006年到2010年，公司主攻Windows，在Linux上反而有回落；同时开始进行RCS的研发与Exploit漏洞利用的处理。

在2011年到2012年，开始向各个操作系统伸出触角，BlackBerry、Symbian、Android等手机操作系统为其主攻方向；而在网络攻击方面则继续在RCS及Exploit业务上努力，RCS更是作为主要业务不断增大影响力。

而2013年到2015年，公司开始了全方位的发展，主要在Android系统进行服务，也恰巧见证了Symbian的衰落。