2016 年中国可视化与可视分析大会

数据可视分析挑战赛-挑战 2

（ChinaVis Data Challenge 2016 - mini challenge 2）

答　卷

参赛队名称：　　中国科学院信息工程研究所-田甜

团队成员：　田甜，中国科学院信息工程研究所，tina_sweet877@sina.com，队长

陈璐，中国科学院信息工程研究所，chenlu. seu@gmail.com

刘松，中国科学院信息工程研究所，lstp678@yeah.net

汪鑫，中国科学院信息工程研究所，wangxin32@126.com

刘俊荣，中国科学院信息工程研究所，ljr_0527@163.com，指导老师

是否学生队（是或否）：　是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，Gephi，Echarts

共计耗费时间（人天）：　30 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

（灰色字为参赛信息填写模板，请参赛者在提交时参照模板填写）

挑战 2.1： 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。（请将回答尽量控制在 2000 个字和 10 张图片内）

HT 人员名单的确定：

如果一个账户多次使用 hackingteam 的域名进行发收邮件我们初步认为其是 HT 成员。得到初始名单后进一步对其进行筛选。设定如果某人发送的邮件数等于 0 并且收发邮件的数量小于 500 则不是有效的 HT 人员。经过筛选最终得到了 132 名 HT 成员列表。利用 Gephi 生成员工关联关系拓扑图如图 1 所示。为了方便、快捷的对各个维度的进行统计，采用 Spark 对邮件进行各种数据处理与统计，Spark 提供了丰富的算子可以完成各种各样的统计任务，并且通过 Spark GraphX 构建整个人员沟通拓扑图基本信息，从而方便的了解每一个人的各种信息。

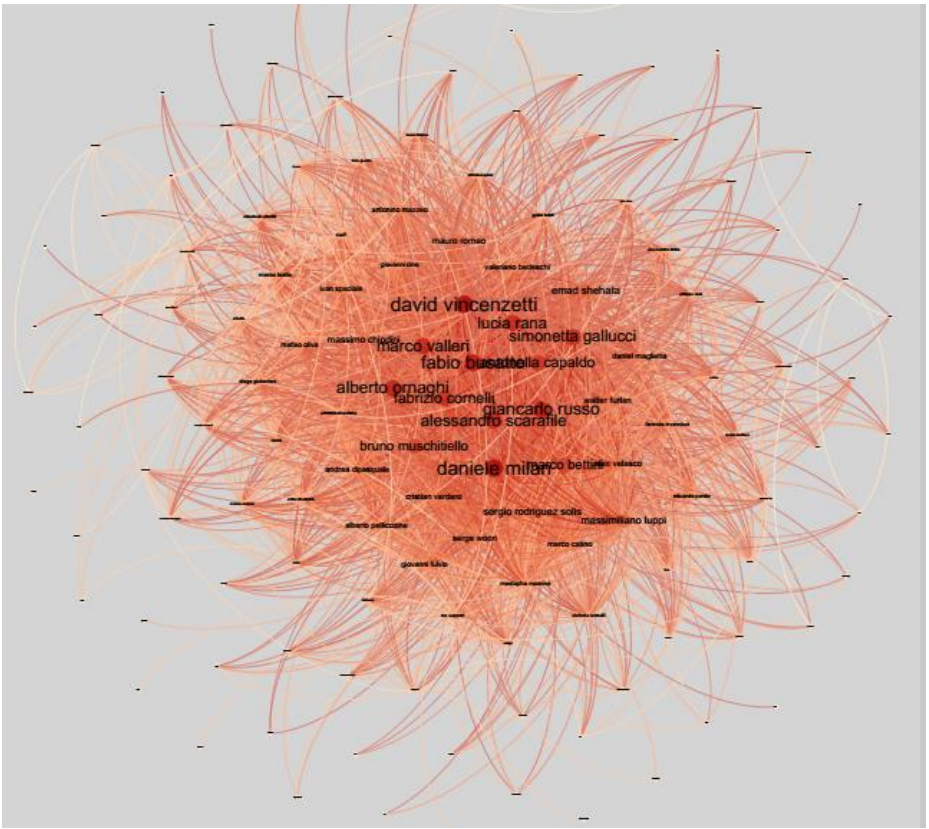

图 1 Hacking team 人员关联关系拓扑图

对于员工的分类，采用了两种分类思路：

1）员工在公司的重要程度

2）通过公司的业务对员工进行分类

1. 员工在公司的重要程度

首先我们对 Hacking Team 的成员依据其在公司的重要程度做分类，重要程度的评判标准选取了四个维度：该员工收到的重要程度为 2 的邮件数量、邮件关联人数的多少、该员工的入职时间以及该员工收发邮件的总量。如果这四个维度都显示该成员有较高的重要程度的话，那么我们就可以认为该成员级别比较高，如图 2 所示。

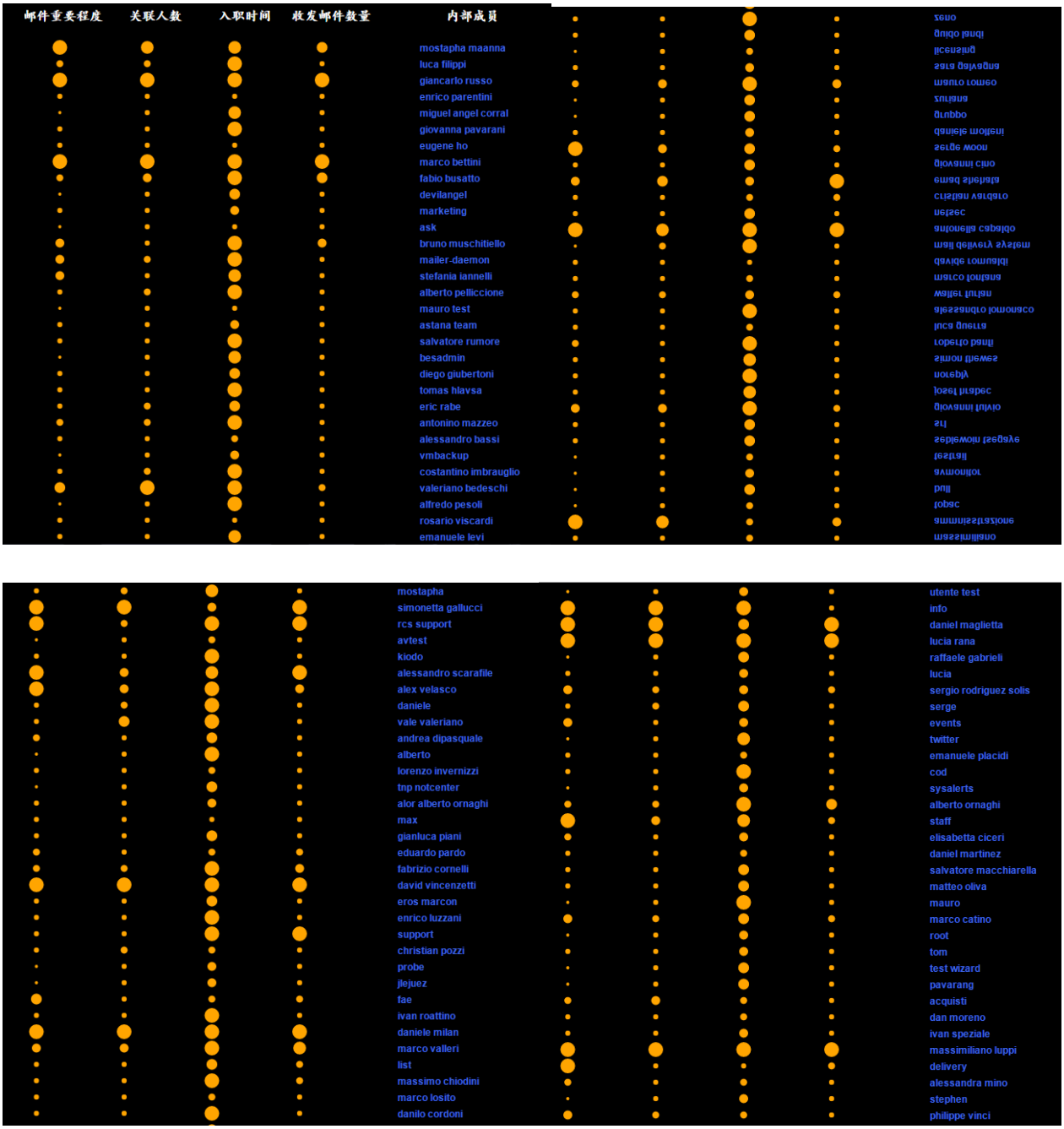

图 2 人员重要程度可视化展示

由图 2 可以发现，四个维度的权值都比较高的员工，得出下面的这些人在公司很重要：David Vincenzetti、Marco Valleri、Giancarlo Russo、Marco Bettini、Alessandra Mino、Daniele Milan、Serge Woon、Daniel Maglietta、Massimiliano Luppi、Giancarlo Russo、 Alberto

Ornaghi、Diego Giubertoni、Enrico Parentini、Fabio busatto、Alessandro Scarafile、Fabrizio Cornelli、Emad Shehata。

为了验证这个可视化效果得出的人物名单，我们对于每个人员的行为特点做了可视化展示，展示内容为：

1）员工在 2011 年-2015 年每小时发邮件数量，该统计情况以散点图的形式展示

2）用折线图来可视化展示该成员在这五年的时间与域名邮件数量的趋势统计

我们通过人员重要程度的散点图可视化展示方式分析出公司内部的重要成员，同时支持多视图可视分析功能。如图 3 所示，对 David Vincenzetti 做进一步分析，由环状散点图可以得出，David Vincenzetti 的主要发邮件时间在早上 9 点到晚上 2 点，而且随着年份的增加，发邮件的数量也比以往更多（其中圆圈的大小表示发邮件数量的多少），这也说明随着时间的发展，该人员的业务量也在不断地加大。通过最右边的折线图，我们能够清楚的了解到，与该人员有业务往来的公司以及邮件来往数量趋势。

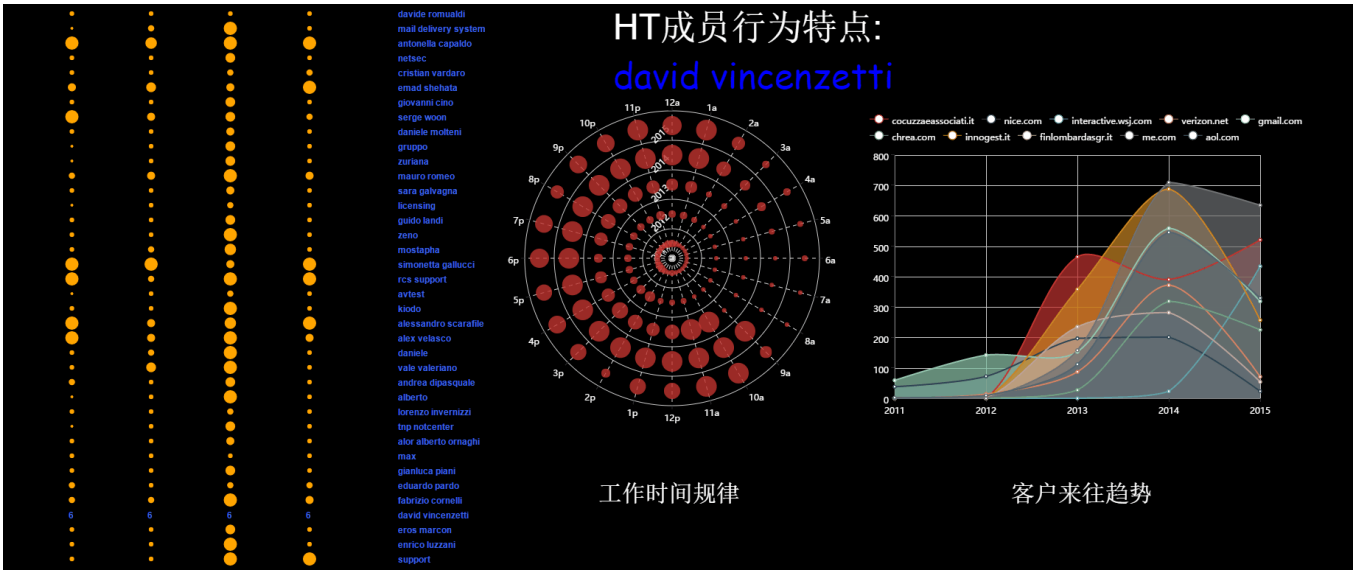

图 3 人员行为特点分析

2. 从业务角度对人员分类

采用了概念图的方式，从工作职责的角度对公司内部成员做了分类，如图 4 所示，通过对数据的分析与提取，获得了该公司业务基于操作系统的业务分类方向：windows、linux、mac、ios、windows phone、symbian、blackberry 和 andriod。

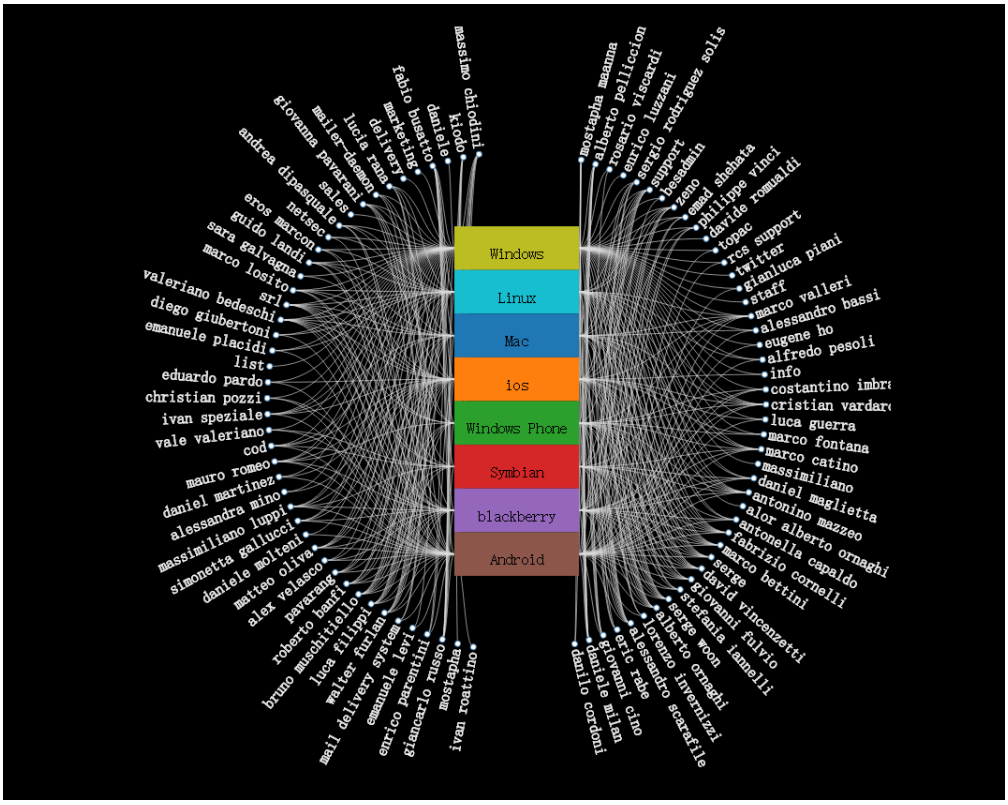

图 4 业务与人员关联图

图 4 的中间矩形框内表示公司所涉及到的业务，四周均分分布的圆圈表示公司的所有内部成员，将光标放在某业务上时，高亮的显示该业务所涉及到的成员。同时，当鼠标光标移动到某 Hacking Team 成员名字上时，也会高亮显示该成员这几年都接触过哪些业务。通过这样的方式我们对人员从业务的角度做了分类，也能通过个人的角度来了解他的业务范畴。例如，图 5 所示为负责 symbian 的人员有：fabio busatto、giovanna pavarani、eros marcon、valeriano bedeschi、vale caleriano、 cod、massimilliano luppi、pavarang、bruno muschitiello、giancarlo russo、alberto prlliccio 等。而图 6 则从个人角度来看他的业务范畴，比如 Marco fontana 这个人的业务有：windows、ios、windows phone。

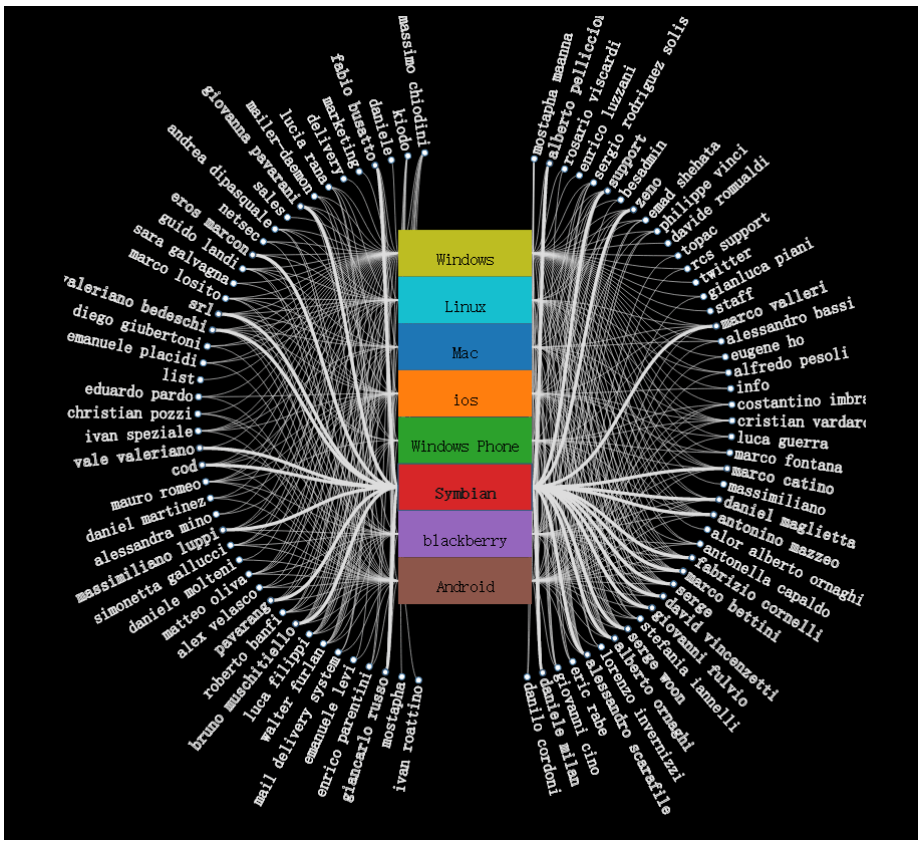

图 5 symbian 涉及到的人员

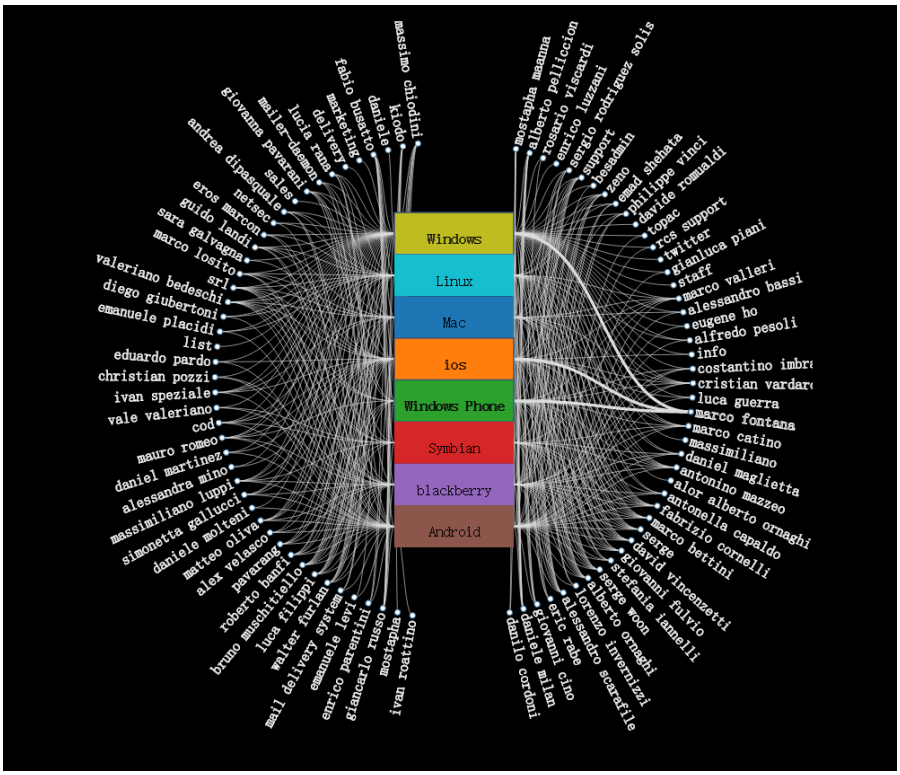

图 6 Marco fontnana 的业务范畴

挑战 2.2： 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。（请将回答尽量控制在 2000 个字和 10 张图片内）

（下面是挑战 2.2 的答题区域）

对于邮件的分类首先分为：业务邮件及非业务邮件。业务邮件包含：操作系统和攻击类型这两类，非业务邮件分析过程中发现主要是生活相关邮件。

基于操作系统的分类有 Windows、Linux、Mac、ios、Windows Phone、Symbian、blackberry、Android 八类；

基于攻击类型的分类有 Exploit、Rcs、Botnet、Malware、0day、DDOS 六类；

基于生活相关的特征主要有 Biglietti-背包客栈、Jvg-乐队、Itinerary-旅程表、Aerei-航空公司、delta-航空公司、Pasticcini-糕点、hotel-宾馆、Anons-广告、Pranzo-意大利风味餐馆、gift-礼物、maglietta -T 恤、ticket、torta-馅饼、visa-银行卡等十五类；

如图 7 所示，分析数据得到的非业务邮件主要是生活相关的主题，包括：订票、旅行、吃饭等主题。

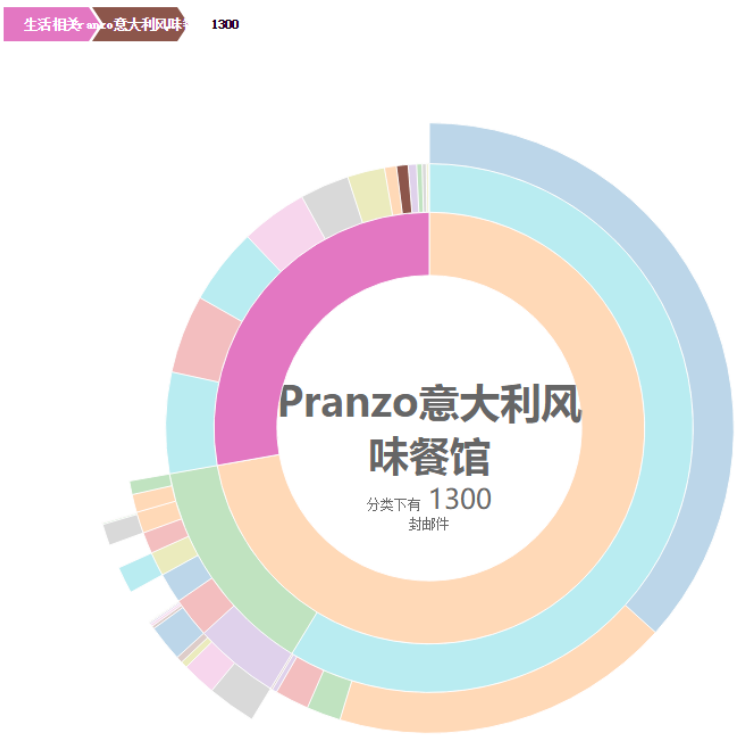

图 7 非业务邮件分类

业务方面，又从操作系统相关和攻击类型两个独立的维度考虑，一个是描述 Hacking team 的产品本身的功能，另一个则描述它所针对的对象。

由图 8 的 sunburst 图，我们可以发现，往来的邮件中，业务邮件占据了近四分之三，其中又有近六分之五的比例是在描述攻击类型（方法）。包括 Exploit，RCS，Botnet，Malware，0day，DDOS 这些常见的黑客手法。而 Exploit 和 RCS 最为显著。RCS 作为 Hacking Team 的最知名的产品，自然频繁出现在邮件中，包括商务洽谈邮件及开发邮件。

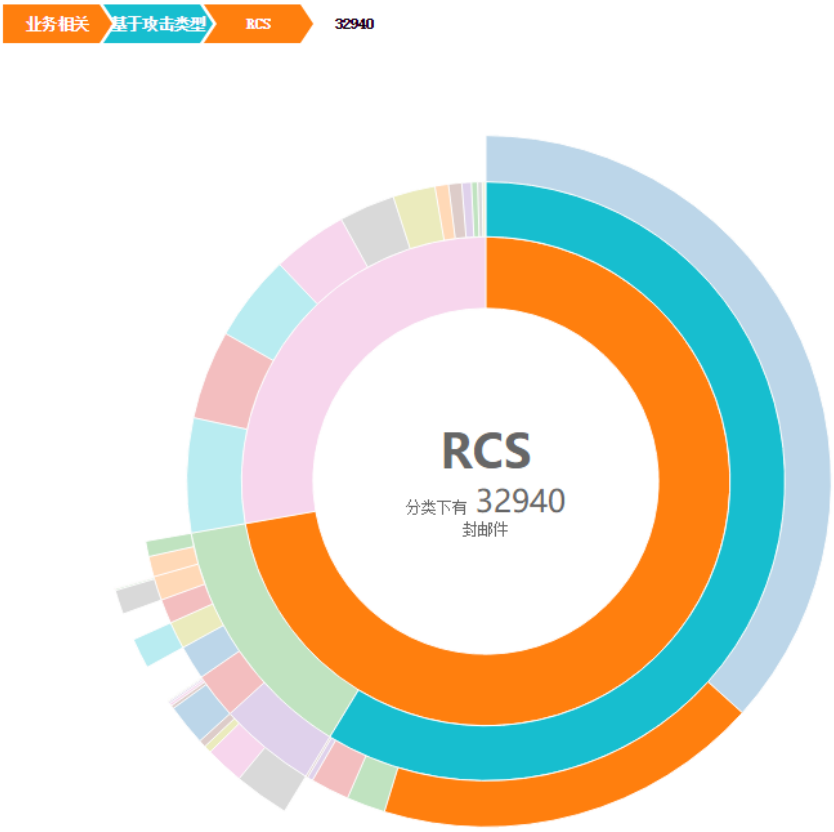

图 8 邮件分类-rcs 层

从操作系统角度，可以看出 Hacking Team 的产品几乎覆盖了所有的 PC 端和移动端，从 android 到 winphone，再到 window，linux 等。均有涉猎。

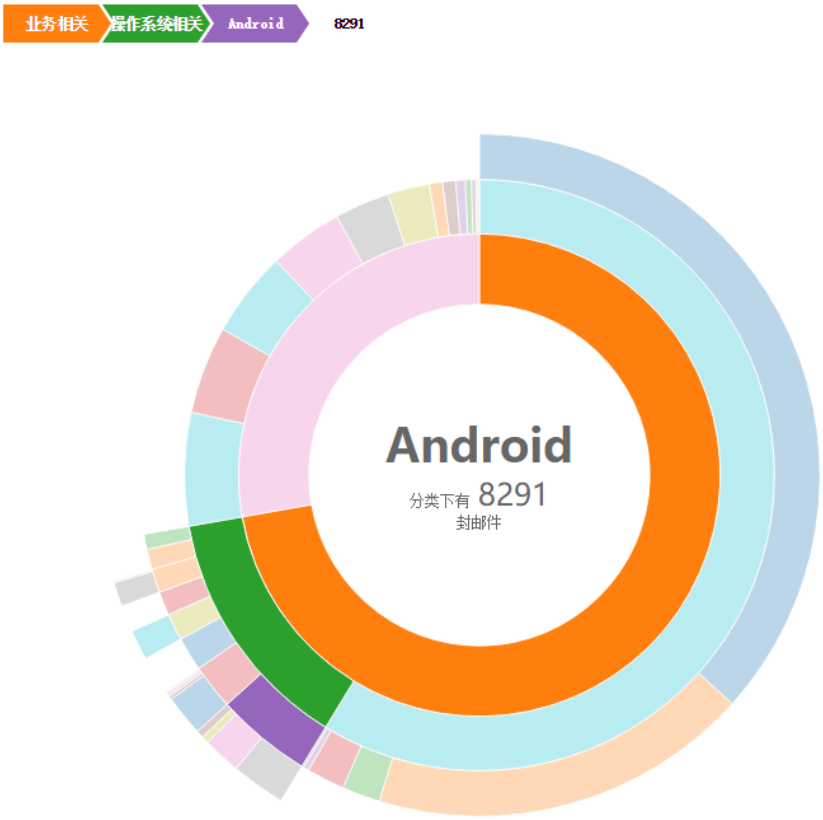

图 9 邮件分类-android 层

挑战 2.3： 根据邮件数据总结 Hacking Team 公司经历了哪几个发展阶段，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。（请将回答尽量限制在 1500 个字和 8 张图片内）

（下面是挑战 2.3 的答题区域）

首先试验着用折线图和河流图来分别对邮件数量、人员数量及涉及到的国家数量进行趋势分析。效果如图 10 所示。

我们首先采用折线图和河流图来分别对邮件数量、人员数量及涉及关联客户所属的国家数量进行趋势分析。从河流图中可以直观的看到，2013 年的 7 月开始，公司的客户逐渐增多，开始第一个增长高峰期，到 2014 年 5 月有个较小的波动，此后数量几乎稳步提升，直到 2015 年 4-5 月达到峰值，再往后便是邮件泄露事件，各国与 Hacking team 的联系数量急剧减少。与客户国家数量增长同步的是邮件数目增长以及大量的开发人员加入，可以看到 2013 年 7 月、2014 年 5 月和 2015 年 6 月邮件数目的变化趋势同客户量的趋势是同步的。最终确定了公司业务发展的关键节点。分别为 2013 年 7 月、2014 年 5 月和 2015 年 6 月。

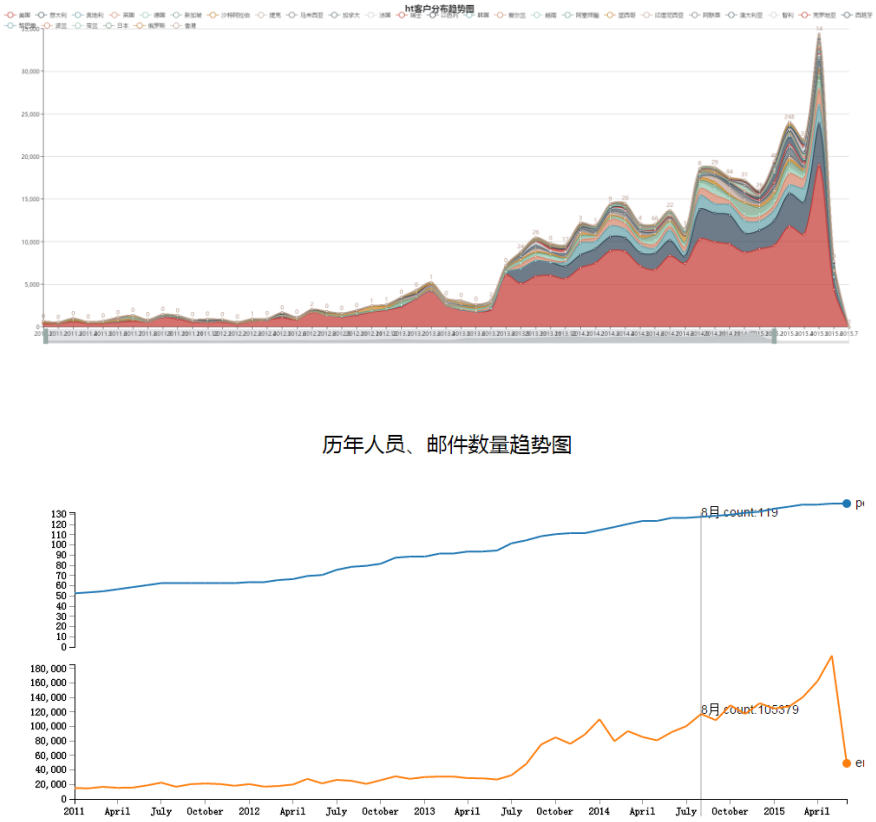

图 10 邮件数量、人员数量及涉及到的国家数量趋势分析统计图

确定了关键节点，下面图 11 对每个节点进行详细分析，用标签云的方式展示从上一个阶段到当前的所有业务及该业务线下的主题词，是一个总的趋势，而右边的折线图则是对业务的一个数量维度的具体展示，是细分到每个月的数据展现。可以看出 12 年 5 月之前 Hacking team 团队的业务以及邮件频繁度都较低，之后开始有起色，从 2013 年 5 月开始 android 便成为公司的主要支撑业务，邮件往来度只增不减。

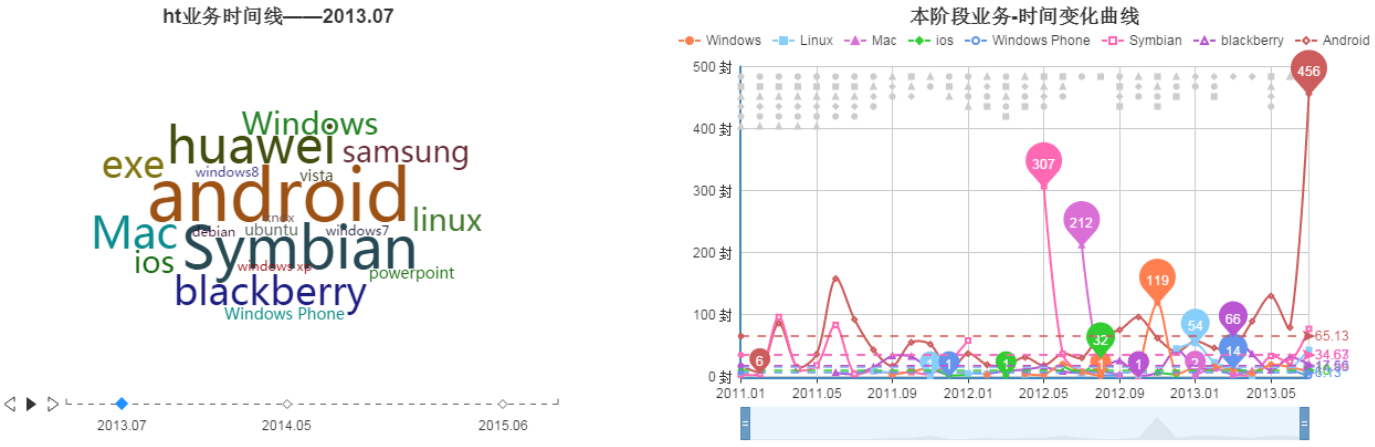

图 11　2013 年 7 月份之前主要业务

由图 11 可以看出 2013 年 5 月份之前这一阶段的主要业务由：Symbian、android、huawei、blackberry、mac、ios、windows、samsung、linux 等。而图 12 所示的 2014 年 5 月份之前的这一阶段的主要业务中 symbian 已经没有什么太多的业务量了，这也符合了当时操作系统的发展趋势，智能手机逐渐占据市场，symbian 渐渐退出操作系统的舞台，samsung 的业务量较上一发展阶段在数量上有着明显的提升。

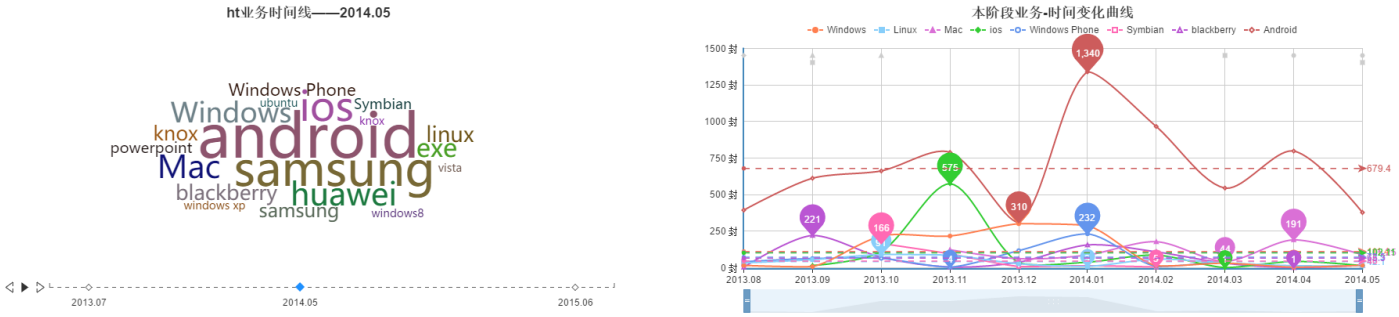

图 12　2013 年 7 月-2014 年 5 月主要业务

而到第二个阶段，2014 年 5 月-2015 年 6 月的时候，如图 13 所示，2014 年 5 月之前 Hackingteam 支持 winxp 的相关业务，在这之后随着微软不在更新 winxp 补丁，客户量骤减，Hackingteam 也停止了 winXP 的业务。2013 年 7 月左右 Android 成为主要支撑业务。直到 2015 年 6 月份，android 的业务发展趋势和客户量趋势及邮件数目趋势基本吻合。从 Android 的业务的发展趋势可视化结果进一步验证了利用折线图和河流图推断公司主要发展阶段时间节点的正确性。

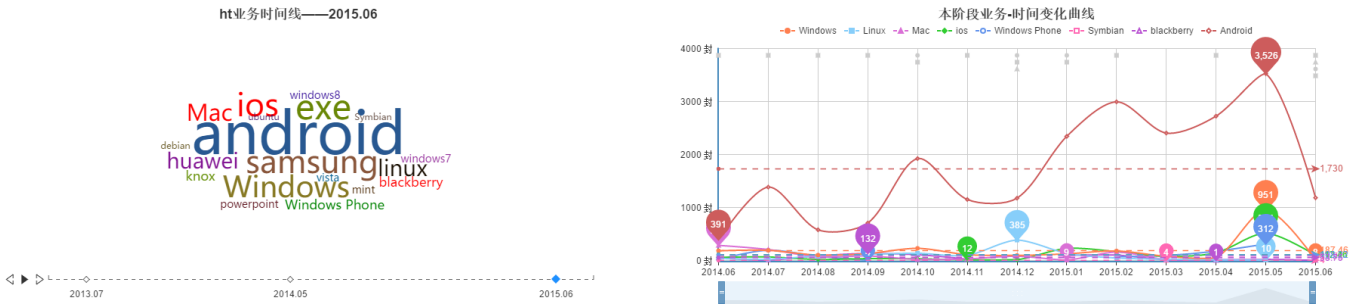

图 13　2014 年 5 月-2015 年 6 月主要业务