**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2**

**背景介绍**

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帐号遭到不明人士入侵，入侵者用此帐号公布了该公司许多内幕信息并通告该公司的内部数据已经泄露。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…”，并附有近400G数据的下载链接，其中包含入侵者所称的内部电子邮件、各种相关文件和源代码。

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。Hacking Team公司内部电子邮件数据是了解该公司的重要数据源，它不但能反映公司员工间及与业务对象间的复杂通信网络，还可以从邮件内容与附件中了解公司经营的业务内容。我们对原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。

**数据说明**

本次提供的邮件数据，数据格式为csv，数据大小为433MB（未压缩），包含61个csv文件，15个字段，共计约120万行行记录。字段及含义见下表：

|  |  |
| --- | --- |
| **字段名称** | **字段含义** |
| Subject | 邮件主题 |
| From(display) | 发件人名称 |
| From(address) | 发件人地址 |
| To(display) | 收件人名称 |
| To(address) | 收件人地址 |
| Cc(display) | 抄送人名称 |
| Cc(address) | 抄送人地址 |
| Bcc(display) | 密送人名称 |
| Bcc(address) | 密送人地址 |
| Creator Name | 创建者名称 |
| Importance | 重要性权值0-Low，1-Normal，2-High |
| Date Sent | 发送时间 |
| Date Received | 接收时间 |
| Size | 邮件大小 |
| Attachment Names | 附件名称 |

下面我们举例说明：

(1). 每个csv文件是与某人相关的所有邮件记录。以a.bassi.csv文件为例，文件中包括Alessandro Bassi所有发送和接收的邮件，其中接收的邮件包括抄送、密送、以及其所在邮件组接收到的所有邮件。

(2). 下表是一条真实的邮件数据，发件人David Vincenzetti将一封主题为“Re: Retribuzioni Maggio”的邮件发送给Simonetta Gallucci，并抄送给CGiancarlo Russo。表中密送人名称与地址字段没有内容，表示这两个字段信息为空。如果字段中包括多个邮件地址、人名或文件名等，利用“；”分割，如下表的Attachment Names字段中包含两个文件Emad Shehata.asc和HT\_NDA\_1.1.pdf，字段值表示为“Emad Shehata.asc;HT\_NDA\_1.1.pdf”。

|  |  |
| --- | --- |
| **字段** | **取值** |
| Subject | Re: Retribuzioni Maggio |
| From(display) | David Vincenzetti |
| From(address) | d.vincenzetti@hackingteam.com |
| To(display) | Simonetta Gallucci |
| To(address) | /O=HACKINGTEAM/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=Simonetta Gallucci569 |
| Cc(display) | Giancarlo Russo |
| Cc(address) | /O=HACKINGTEAM/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=Giancarlo Russof7a |
| Bcc(display) |  |
| Bcc(address) |  |
| Creator Name | David Vincenzetti |
| Importance | 1 |
| Date Sent | 2015/6/5 12:18 |
| Date Received | 2015/6/5 12:18 |
| Size | 14487 |
| Attachment Names | Emad Shehata.asc;HT\_NDA\_1.1.pdf |

(3). From/To/Cc/Bcc (display)中可能出现完整的邮件地址，也可能出现员工或是邮件组的名称甚至昵称。例如上表中From(display)字段的取值“David Vincenzetti”应该是一个员工的姓名。假设From(display)字段的取值为“DV”，那么它可能代表“David Vincenzetti”的一个昵称，也可能代表一个邮件组的名称或昵称。

(4). From/To/Cc/Bcc (address)中可能出现一些很长的邮件地址，这是由于EXCHANGE邮件服务器解析问题导致的。例如上表中的Cc(address)字段的取值， 其中HACKINGTEAM代表Hacking Team公司，Giancarlo Russof7a代表Giancarlo Russo这个人的一个邮件地址在邮件系统中的唯一标识(ID)，此处Giancarlo Russof7a由名字（Giancarlo Russo）+后缀（7a）组成，其它文字信息均为系统保留字。

**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。