安全知识图谱应用—黑灰产团伙挖掘

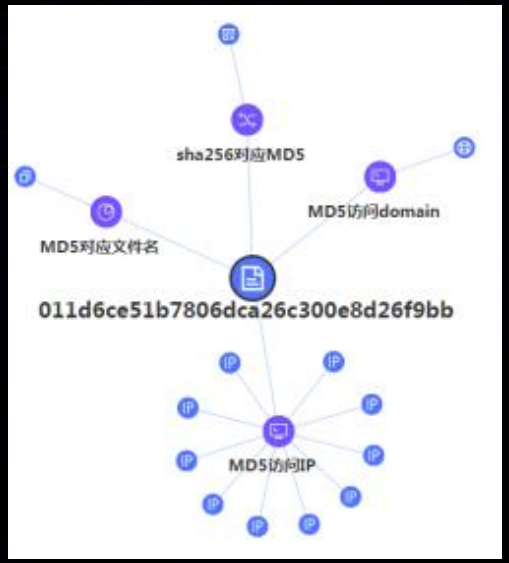
邓永

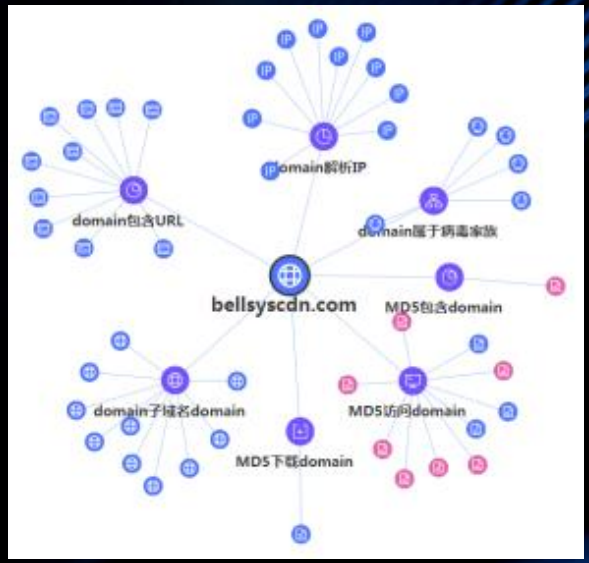
腾讯安全高级研究员

目录

安全图谱介绍——异构图图数据库

- 包含多种安全实体——文件、域名、IP、URL、域名注册信息、病毒家族、漏洞CVE等
- 包含多种实体与实体之间的关系
- 利用学习模型学习的实体与实体之间的关系
  - 域名相似度
  - 文件相似度
- 图数据库能够做到快速深度优先遍历





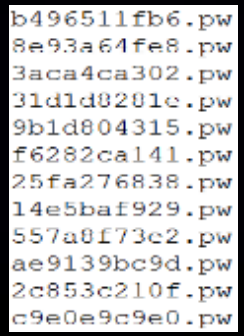
团伙介绍——商贸信家族团伙

安全图谱的应用——黑灰产团伙挖掘

- 团伙的安全实体在图谱中具有很强的**内聚性**
  - 同团伙的实体与实体之间的网络和行为关系连接**紧密**
  - 不同团伙间的实体与实体连接比较**稀疏**
- 利用图聚类算法发现可疑团伙
  - 能够从整体全面分析团伙的行为

b496511fb6.pw

点



面

体

安全图谱的应用——黑灰产团伙挖掘步骤

- 构建图（200+亿节点，1500+亿边，图太大）
  - 直接关系
  - 间接关系
- 图聚类算法（发现可疑类簇）
  - FastUnfolding算法
- 聚类结果分析
  - 基于规则和人工运营的方法

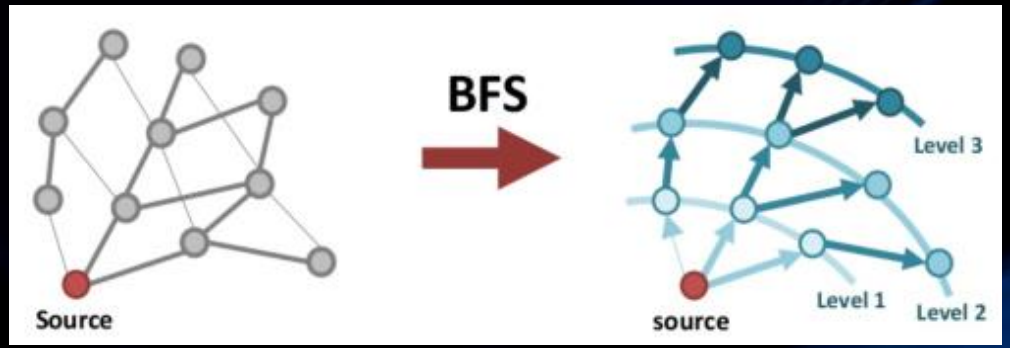
构建实体与实体关系图

- **直接关系：实体网络关系和实体行为关系**
  - 直接关系能够通过日志、沙箱和外部数据直接得到的关系对；
  - 例如文件访问域名、域名解析IP、文件访问IP、域名与注册者等关系

- **间接关系：通过非图谱中的实体间接关联而学习得到**
  - 文件与文件相似度关系
  - 域名与域名相似度关系

构建实体与实体关系图

- **直接关系：实体网络关系和行为关系**
  - 图数据库中通过广度优先遍历
  - 遍历过程中对数据进行清洗和降噪
  - 起始节点的选择问题



构建实体与实体关系图

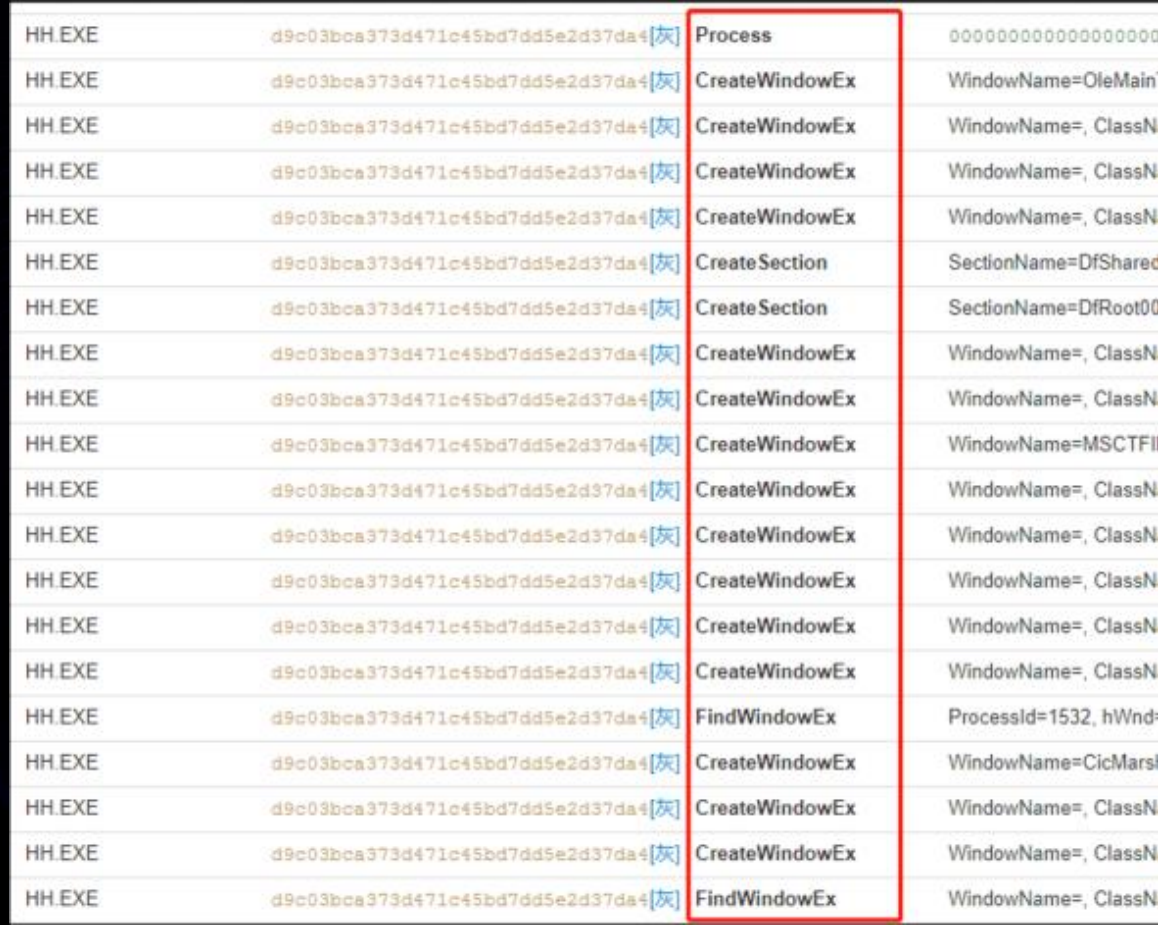
- **间接关系：通过非图谱中的实体间接关联而学习得到**
  - **文件与文件相似度关系**
    - 文件与文件通过**静态、动态和网络**行为特征建立关联
    - 文件与文件通过**uid**建立关联
  - **域名与域名相似度关系**
    - 域名与域名通过**uid**建立关联

构建图——文件相似性

- **文件静态相似性**
- 文件动态相似性
- 文件网络行为相似性

构建图——文件相似性

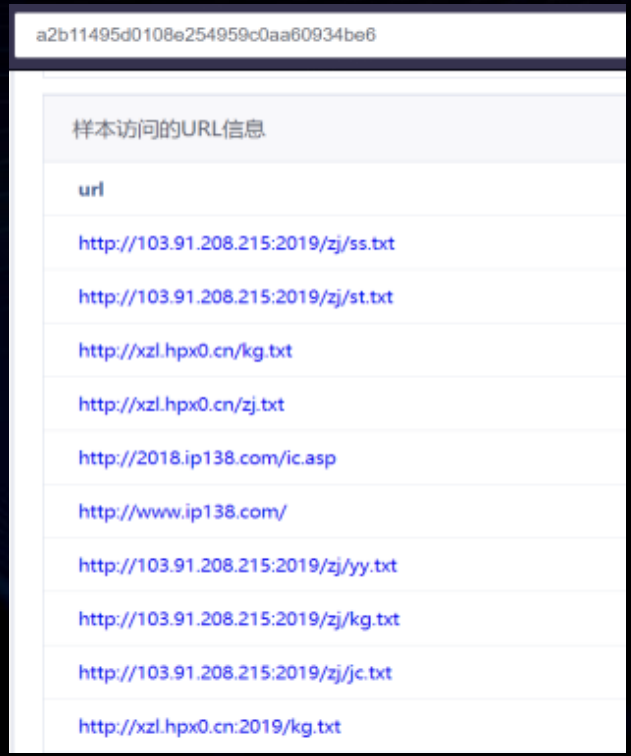
- 文件静态相似性
- **文件动态相似性**
- 文件网络行为相似性

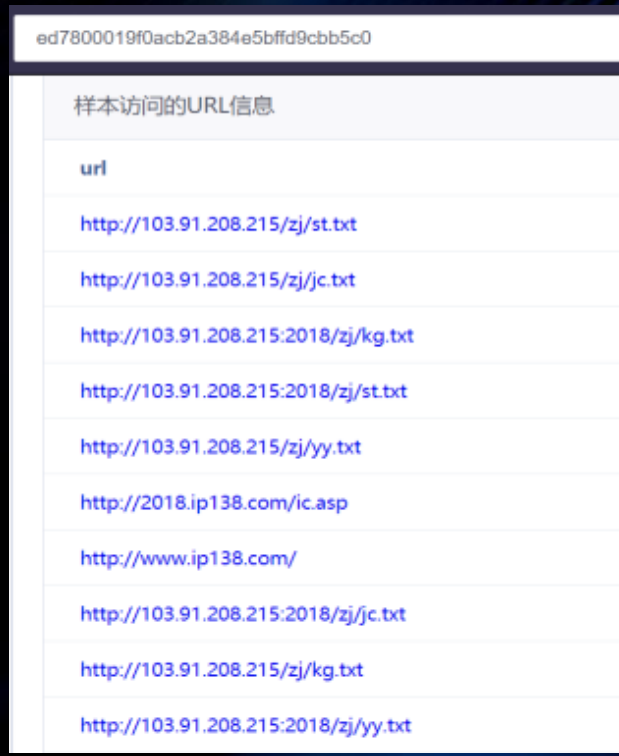


HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	Process	0000000000000000
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=OleMain
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateSection	SectionName=DfShared
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateSection	SectionName=DfRoot00
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=MSCTFI
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	FindWindowEx	ProcessId=1532, hWnd=
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=CicMars
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	CreateWindowEx	WindowName=, ClassN
HH.EXE	d9c03bca373d471c45bd7dd5e2d37da4[灰]	FindWindowEx	WindowName=, ClassN

构建图——文件相似性

- 文件静态相似性
- 文件动态相似性
- **文件网络行为相似性**





构建图——基于simhash计算文件相似性

SimHash是一种局部敏感hash。它也是Google公司进行海量网页去重使用的主要算法。它通过将原始的文本映射为二进制数字串，然后通过比较二进制数字串的差异进而来表示原始文本内容的差异。

你妈妈喊你回家吃饭了

你妈妈叫你回家吃饭啦

通过传统hash计算为：

00010000011001101001110110111 10

10100100011111111100101100111 01

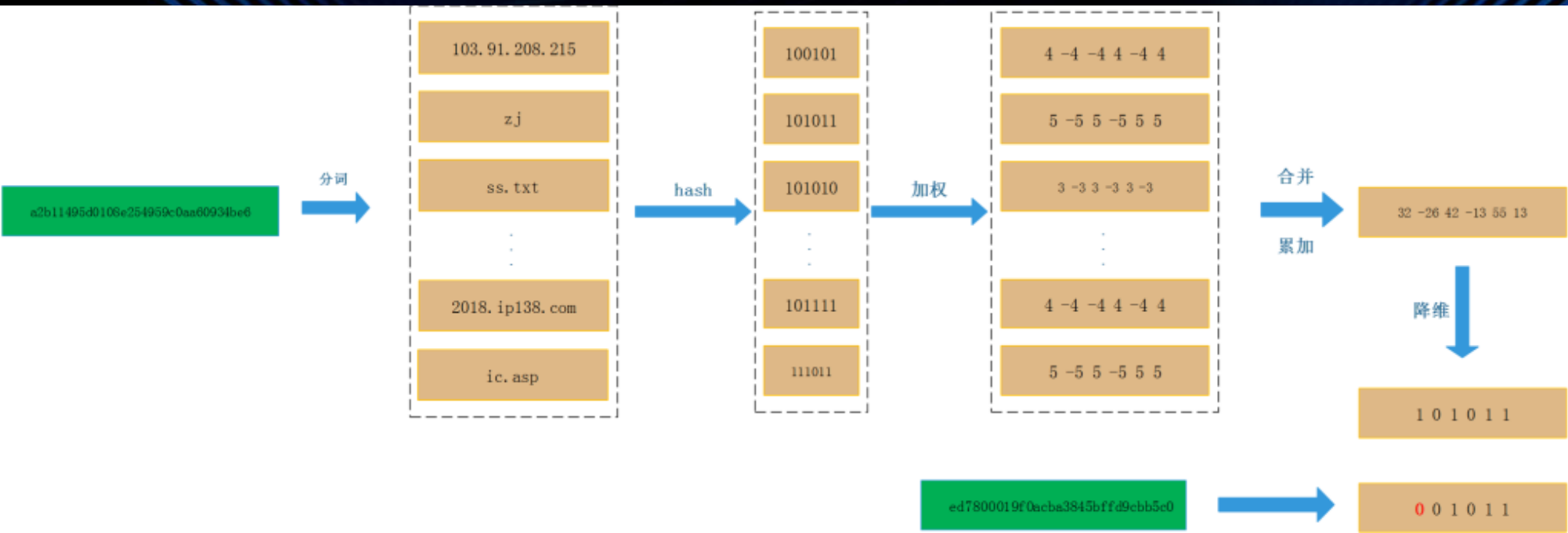
通过simhash计算结果为：

10000100101011011111110000010101101000100111110000100101100101 1

10000100101011010111110000010101101000100111110000110101000101 1

构建图——文件相似性

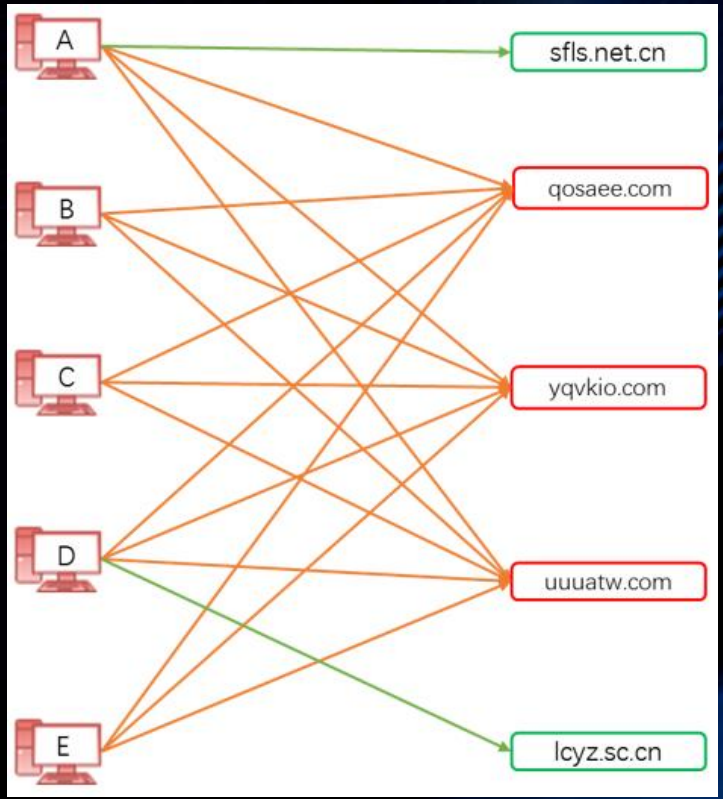
- **Simhash计算文件网络行为相似性**



构建图——域名似性

- 如果域名属于相同恶意家族，存在大量相同的主机访问它们，访问主机具有较高重合度

- 访问主机重合度越高域名属于相同团伙的概率越大



构建图——域名似性

- 如果域名属于相同恶意家族，存在大量相同的主机访问它们，访问主机具有较高重合度
- 访问主机重合度越高域名属于相同团伙的概率越大
  - 主机集合的Jaccard相似度越大为相同团伙的概率越高
  - 两两计算域名相似度时间复杂度平方级

$$\mathrm{Sim}(d_1, d_2) = \mathrm{Jaccard}(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

访问域名$qosaee.com$的主机集合

S_1={A, B, C}

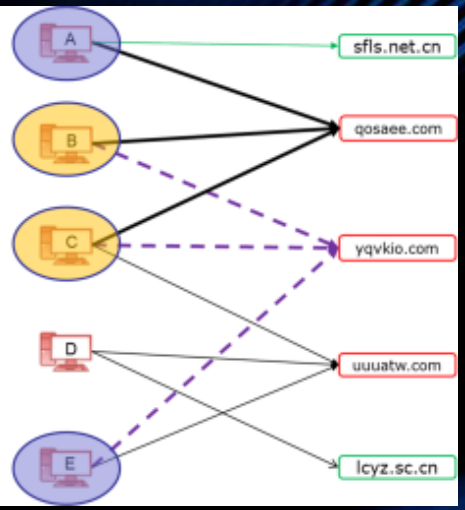
访问域名yqvkio.com的主机集合

S_2={B, C, E}

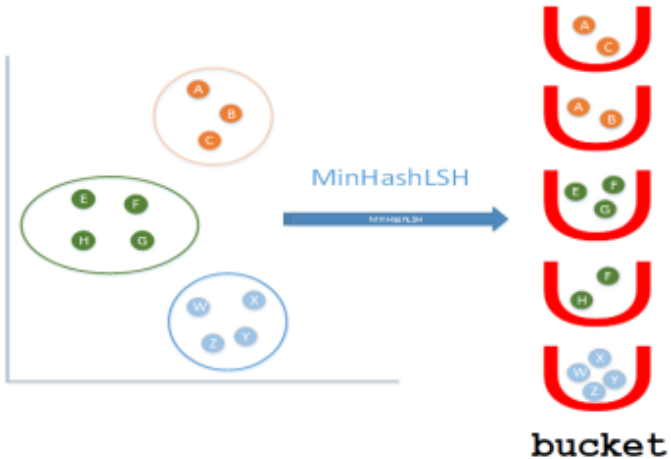
$$\mathrm{Sim}(\mathrm{qosaee.com}, \mathrm{yqvkio.com}) = \mathrm{Jaccard}(S_1, S_2)$$

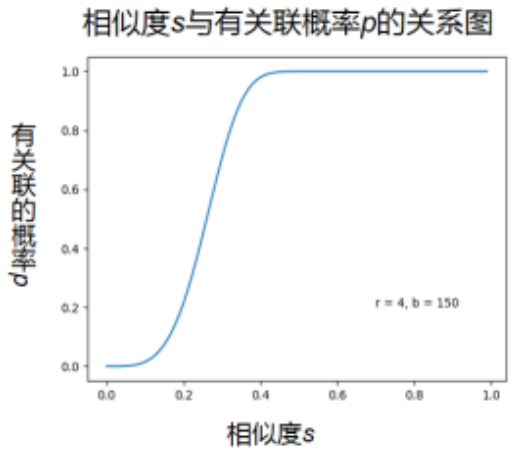
$$= \frac{|\{B,C\}|}{|\{A,B,C,E\}|} = \frac{1}{2}$$

域名相似度越高，它们为相同家族的概率越大



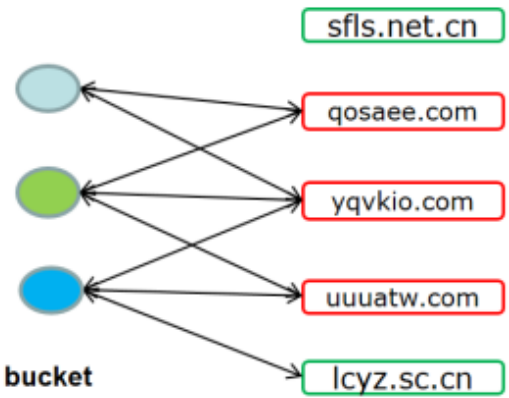
构建图——基于MinHashLSH计算域名似性





如果两个域名的Jaccard相似度s大于给定阈值t，则它们以较大概率p至少映射到一个相同bucket

- 相似度大的域名通过bucket关联（**同家族高概率关联**）
- 相似度小的域名无关联（**不同家族低概率关联**）



图聚类——FastUnfolding算法

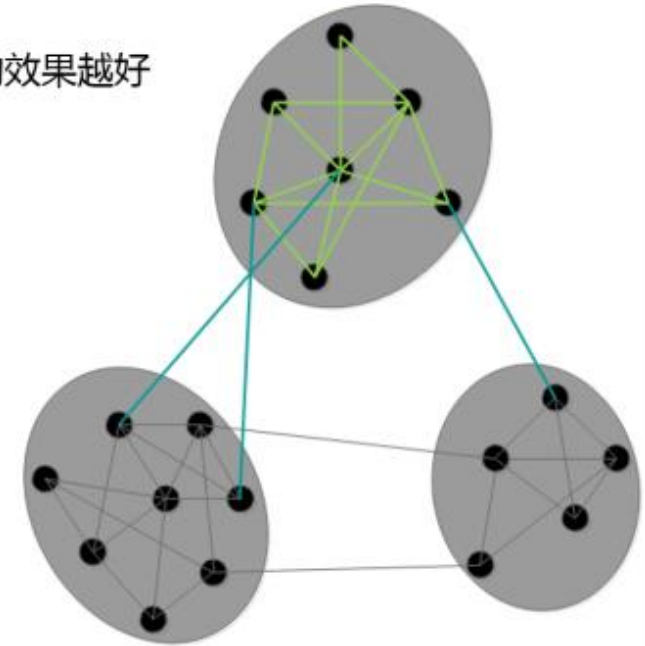
模块度(modularity)

衡量社区划分的优劣，模块度越大，则社区划分的效果越好

社区内部所有边的权重合

$$Q = \sum_{c \in C} \left[\frac{\Sigma_{in}^{c}}{2m} - \left(\frac{\Sigma_{tot}^{c}}{2m} \right)^{2} \right]$$

社区与外部相关联的边权重各



图聚类——FastUnfolding算法

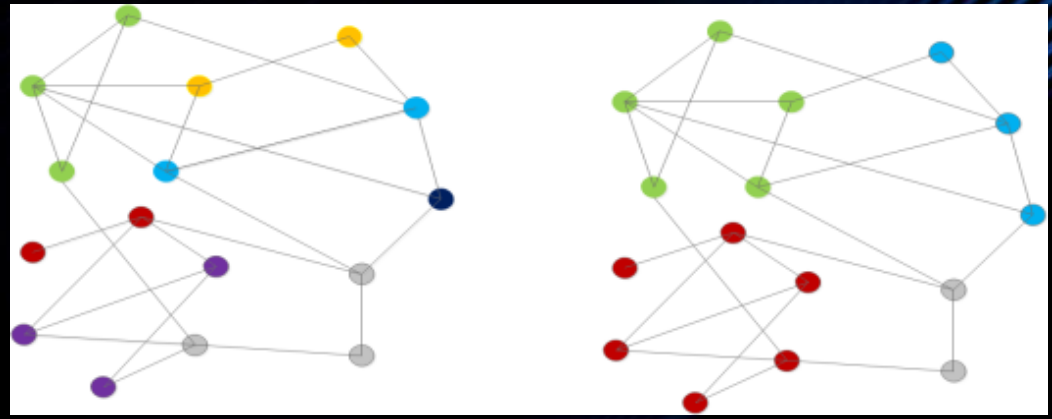
模块度(modularity)

衡量社区划分的优劣，模块度越大，则社区划分的效果越好

社区内部所有边的权重合

$$Q = \sum_{c \in C} \left[\frac{\sum_{in}^{c}}{2m} - \left(\frac{\sum_{tot}^{c}}{2m} \right)^2 \right]$$

社区与外部相关联的边权重各



Q1 < Q2

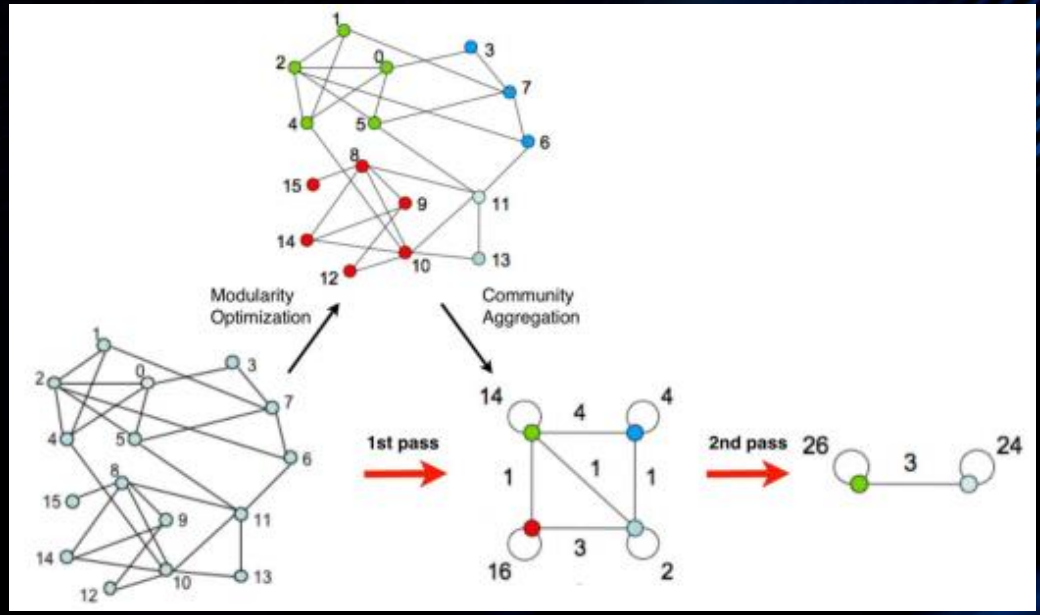
图聚类——FastUnfolding算法

第一阶段：模块度优化

将网络中的每个节点看成一个独立的社区，然后不断地遍历网络中的结点，尝试将单个结点加入能够使模块度Q提升最大的社区中，直到所有结点都不再变化。

第二阶段：图折叠

处理第一阶段的结果，将一个个小的社区归并为一个超结点来重新构造网络，这时边的权重为两个结点内所有原始结点的边权重之和。迭代这两个步骤直到算法稳定。



参考：Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

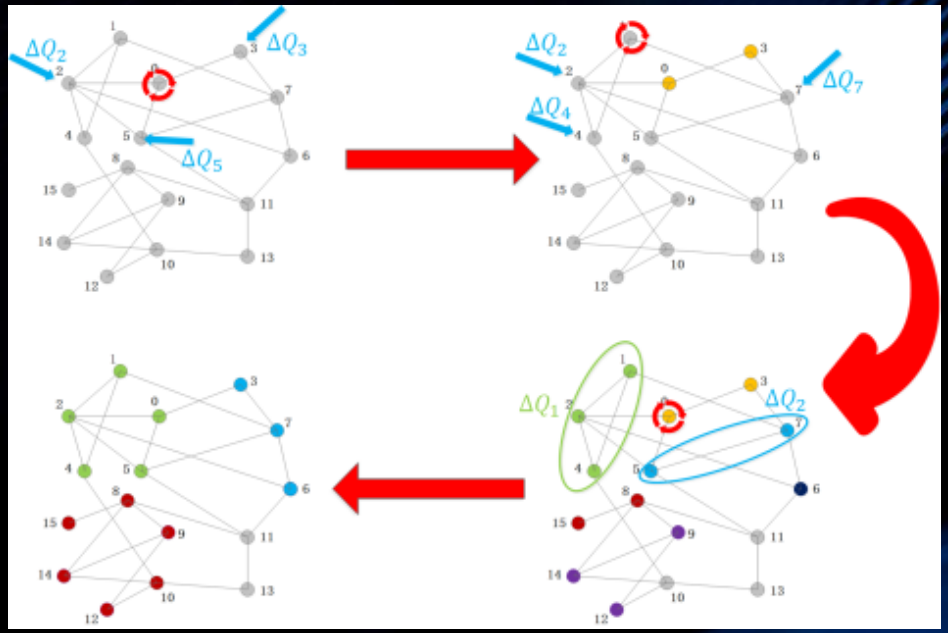
图聚类——FastUnfolding算法

第一阶段：模块度优化

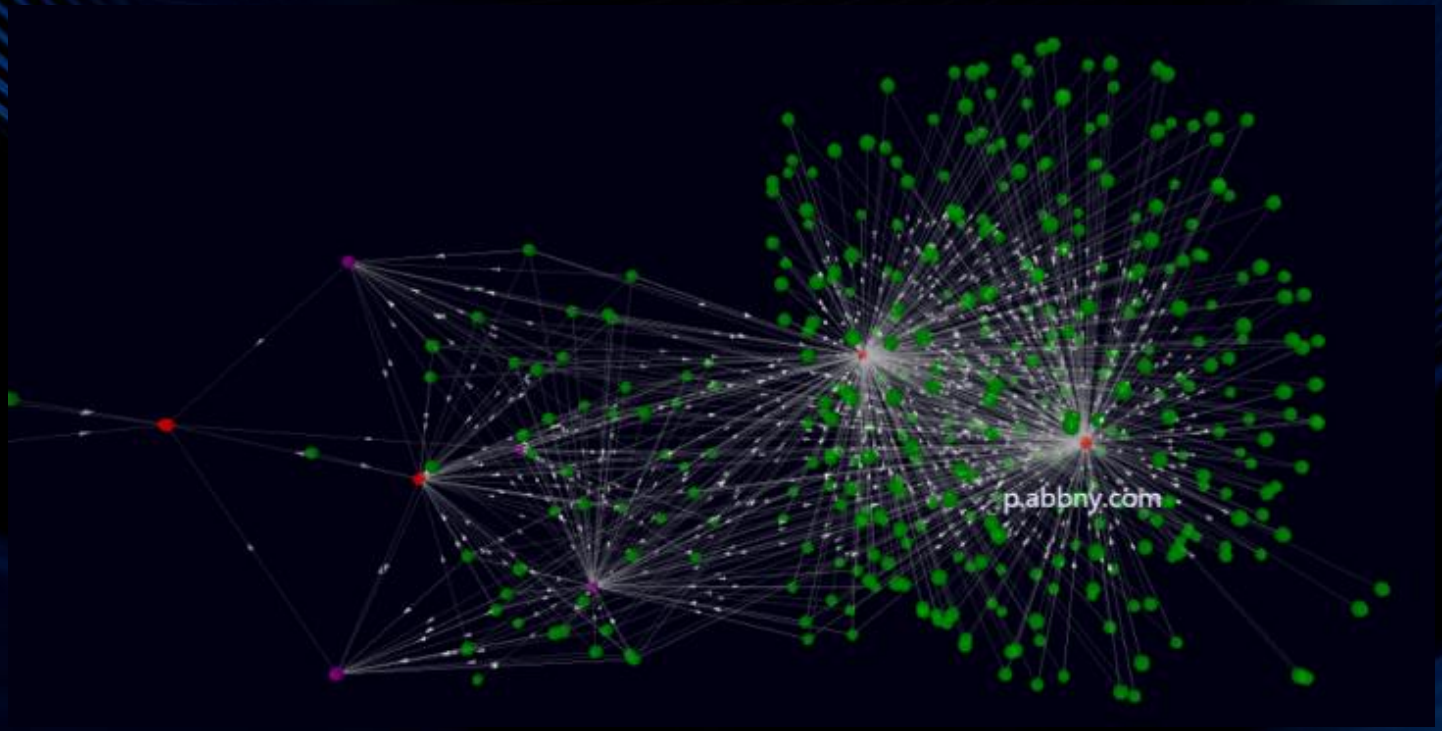
将网络中的每个节点看成一个独立的社区，然后不断地遍历网络中的结点，尝试将单个结点加入能够使模块度Q提升最大的社区中，直到所有结点都不再变化。

第二阶段：图折叠

处理第一阶段的结果，将一个个小的社区归并为一个超结点来重新构造网络，这时边的权重为两个结点内所有原始结点的边权重之和。迭代这两个步骤直到算法稳定。



团伙聚类结果举例——“驱动人生”木马团伙



团伙聚类结果举例——“抓鸡狂魔” 团伙

安全图谱应用总结及未来展望

- **构建图**
  - 同类连接紧密
  - 非同类连接稀疏
- **图聚类算法（发现可疑类簇）**
- **可疑类簇分析判定**
  - **判定聚类为恶意团伙**
  - **判定类簇具体行为**


THANKS
TENCENT SECURITY CONFERENCE 2019