## 计算Jaccard系数

为了计算恶意软件样本之间的Jaccard系数，我们首先通过提取我们想要使用的特征来估计样本共享的代码量。这些可以是前面描述的任何特征，例如基于导人地址表的函数、字符串、N-gram指令或者是N-gram动态行为。在这里，我们将使用可打印字符串作为特征，原因是它们的执行效果很好并且易于提取和理解。

以下分别是我们提取的三个恶意软件样本的特征集合构成的特征袋：

**样本1“F6655E39465C2FF5B016980D918EA028”：**

{\_\_p\_\_fmode，%X @，WaitForSingleObject，\*(SY)#，=H @，D$ RP，SSPS，\_except\_handler3，DisconnectNamedPipe，GetCurrentProcess，\_XcptFilter，USER32.dll，h !@，D$dSV，=@ @，sprintf，WriteFile，L$(h，=, @，ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-/，%p @，t2Ht，PeekNamedPipe，%d @，T$\h@0@，D$(Qj，.rsrc，MSVCRT.dll，PADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADD，Rich，T$4hT0@，D$DP，%` @，D$ h，D$\QPU，\_beginthread，%\ @，=( @，exit，T$(Q，D$8D，\_acmdln，cmd.exe，LoadStringA，CreateThread，XPVSS，free，\_\_getmainargs，ExitThread，D$\h，\_^][，TerminateProcess，\$Hf，L$(hT0@，@.data，\_exit，t$p3，TerminateThread，SetEvent，KERNEL32.dll，T$4hp0@，%h @，[t V，\_itoa，GetStartupInfoA，L$,h，T$ QRP，WaitForMultipleObjects，SUVWj:h，T$0f，send = %d，\*(SY)# ，atol，GetComputerNameA，DuplicateHandle，\_\_setusermatherr，QRSSSj，UVWj，!This program cannot be run in DOS mode.，CreateEventA，\_strnicmp，\_controlfp，WS2\_32.dll，%l @，T$,h，malloc，hSVW，D$$V，\$lf，.text，L$$VQR，%x @，|$XS，`.rdata，\_\_p\_\_commode，strchr，\_adjust\_fdiv，GetModuleHandleA，ReadFile，L$\hL0@，=8 @，CloseHandle，\*(SY)# cmd，\_initterm，5< @，SVWj，CreatePipe，>"u:F，CreateProcessA，Sleep，L$pj，\_\_set\_app\_type}

**样本2“STARSYPOUND\_sample\_9EA3C16194CE354C244C1B74C46CD92E”：**

{\_\_p\_\_fmode，%X @，WaitForSingleObject，\*(SY)#，=H @，D$ RP，SSPS，\_except\_handler3，DisconnectNamedPipe，GetCurrentProcess，\_XcptFilter，USER32.dll，h !@，D$dSV，=@ @，sprintf，WriteFile，L$(h，=, @，ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-/，%p @，t2Ht，PeekNamedPipe，%d @，T$\h@0@，D$(Qj，.rsrc，MSVCRT.dll，Rich，T$4hT0@，D$DP，%` @，D$ h，D$\QPU，\_beginthread，%\ @，=( @，exit，T$(Q，D$8D，\_acmdln，PAPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDINGPADDINGXXPADDING，cmd.exe，LoadStringA，CreateThread，XPVSS，free，\_\_getmainargs，ExitThread，D$\h，\_^][，TerminateProcess，\$Hf，L$(hT0@，@.data，\_exit，t$p3，TerminateThread，SetEvent，KERNEL32.dll，T$4hp0@，%h @，[t V，\_itoa，GetStartupInfoA，L$,h，T$ QRP，WaitForMultipleObjects，SUVWj:h，T$0f，send = %d，\*(SY)# ，atol，GetComputerNameA，DuplicateHandle，\_\_setusermatherr，QRSSSj，UVWj，!This program cannot be run in DOS mode.，CreateEventA，\_strnicmp，\_controlfp，WS2\_32.dll，%l @，T$,h，malloc，hSVW，D$$V，\$lf，.text，L$$VQR，%x @，|$XS，`.rdata，\_\_p\_\_commode，strchr，\_adjust\_fdiv，GetModuleHandleA，ReadFile，L$\hL0@，=8 @，CloseHandle，\*(SY)# cmd，\_initterm，5< @，SVWj，CreatePipe，>"u:F，CreateProcessA，Sleep，L$pj，\_\_set\_app\_type}

**样本3“WEBC2-AUSOV\_sample\_097B5ABB53A3D84FA9EABDA02FEF9E91”：**

{\_\_p\_\_fmode，vX |iR ，h0A@，\_except\_handler3，%P0@，ExpandEnvironmentStringsA，\_XcptFilter，%X0@，GetTempPathA，h8A@，%temp%，h0@@，-$81# 6%2+:，%x0@，strstr，PhxA@，strncmp，VWh|@@，LZ32.dll，RegCreateKeyExA，GetModuleFileNameA，strchr，2??,，exit，=<0@，strrchr，\_acmdln，GetModuleHandleA，XPVSS，h$@@，\_\_getmainargs，hD@@，t \_^，RegSetValueExA，@.data，\_exit，%\0@，RegCloseKey，vX Rich，LZOpenFileA，hP$@，KERNEL32.dll，MSVCRT.dll，LoadLibraryA，%l0@，strncpy，h A@，GetStartupInfoA，urlmon.dll，InternetCloseHandle，DeleteFileA，\_\_setusermatherr，InternetReadFile，\_\_p\_\_commode，LZCopy，atoi，hh@@，GetProcAddress，hX@@，URLDownloadToFileA，InternetOpenA，!This program cannot be run in DOS mode.，CopyFileA，ADVAPI32.dll，\_controlfp，%L0@，vX |i\ ，hSVW，i]\@GZ.，.text，U^CD，hD$@，`.rdata，\_adjust\_fdiv，%H0@，wininet.dll，\svchost.exe，CloseHandle， -s ，\_initterm，InternetOpenUrlA，>"u:F，CreateProcessA，GetLongPathNameA，Sleep，LZClose，vY ;vX %PS ，\_\_set\_app\_type}

各样本提取的特征数如表1所示。

表1 各样本特征数

|  |  |
| --- | --- |
| 样本名 | 特征数量 |
| F6655E39465C2FF5B016980D918EA028 | 113 |
| STARSYPOUND\_sample\_  9EA3C16194CE354C244C1B74C46CD92E | 113 |
| WEBC2-AUSOV\_sample\_  097B5ABB53A3D84FA9EABDA02FEF9E91 | 89 |

根据Jaccard系数计算公式：Jaccard系数=共享特征数/总特征数，我们分别计算这三个样本之间的Jaccard系数，如表2所示。

表2 各样本对的Jaccard系数

|  |  |  |  |
| --- | --- | --- | --- |
| 样本对 | 共享特征数 | 总特征数 | Jaccard系数 |
| 样本1和样本2 | 112 | 114 | 0.98 |
| 样本1和样本3 | 29 | 173 | 0.17 |
| 样本2和样本3 | 29 | 173 | 0.17 |

各软件样本之间的Jaccard系数代表它们的代码共享程度，我们通过绘制相似性图来可视化这种代码共享关系，如图1所示。

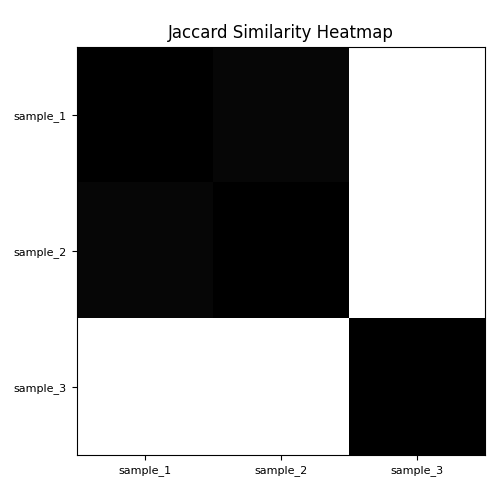


图1 恶意样本相似性图

在图1中，色块颜色的深浅反映了样本间的代码共享程度，颜色越深代表共享程度越高。我们通常把共享程度高于某一阈值的样本归类为同一家族。