Product Annotation Guide

Rebecca S. Portnoff, Sadia Afroz, Greg Durrett, Jonathan K. Kummerfeld,
Taylor Berg-Kirkpatrick, Damon McCoy, Kirill Levchenko, Vern Paxson

UCB, ICSI, NYU, UCSD

1   General Guidelines

Annotate every point in the post when the product is referred to explicitly. In each case, label the single word that best captures its meaning.

- To indicate the word, enclose it in:[1]

    - If it is being sold, use '{' and '}'
    - If it is being bought, use '[' and ']'

  or

    - If it is being sold, use '{S' and '}'
    - If it is being bought, use '{B' and '}'

- Braces should enclose one whitespace delimited token.

- Only annotate cases in the first 10 lines and last 10 lines of a file (not counting blank lines). The numbers at the start of lines provide the relevant count.

- Do not annotate text inside vouches, marked by lines saying ¡blockquote¿ (but vouches do count towards the 10 line rule).

- If there are multiple products, annotate all of them. However, do not annotate features of a single product.
  e.g. do not label *Logo* in *Logo for the [app] 5 $*

- If a product is referred to with a generic word then do not annotate it.
  e.g. *somebody*, *it*, *get*
  The exception to this rule is if the product is never referred to more specifically, then do annotate.
  e.g. *a [method] for making money*

- If the product is a service that produces an item, annotate both.
  e.g. *[hack] a big [website]*, and *I want a coder to [write] an [exploit]*

- Do annotate specific product names if that is the way they are mentioned,
  e.g. *[FaceSpread] // Custom Panel // HTTP crypter and more !*

- Do not annotate unrelated or rival products that are mentioned
  e.g. *similar to Element Scanner or other scanning websites*

---

[1]For this work we do not use the buy/sell distinction that is included in the annotations.

Multi-word expressions Here we are referring to cases like *Remote Administration Tool*, where there are a series of consecutive words forming a single mention of the product. Select only one word, trying to choose the one that best captures the type of product. Some examples of common cases (with recommended annotations) are:

- *Facebook [accounts]*
  Facebook would be less specific, as it could occur with accounts, likes, hacking of the company itself, etc.

- *Phone [Verification] Service*
  Here *Service* would be the next best choice, but *verification* is a better characterization.

- *YouTube Multi Account [Subscriber] Bot*
  Similar idea to the previous case.

One heuristic to try is thinking about what words can and cannot be dropped and still have the product be the same. Those that cannot be dropped are better options for annotation.

1.1 Flags

If a post falls into one of the categories below, we add an extra line at the end of the file and write the letter for that case:[2]

A ADMIN posts, about running the site, deleting a thread, etc

D DIFFICULT cases

W WEIRD cases

G GAMING related posts

L Posts that are not interesting to us (e.g. selling signature space, or shoes)

2 Tricky example cases for reference

Complex products:

- *i am looking to buy USA {Dob} + {ssn} + {fname} + {lname} + {address}*

- *I am buying as many {email} : {passes} as possible.*

- *{Adbot}-{clickbot} {coder} needed.*

- *{WMZ} - [LR]*
  1:1
  Around 200$

Combined items and services:

- *Steal certain @yahoo {email} or facebook {account}*
  i need a certian @yahoo {email} address {hacked} or an associated facebook {account}, paying good $$ if u can do this send me a pm for the info thanx

- *{hack} {website}*
  I pay 2000? via WesternUnion
  PM or add to MSN if interested.

---

[2]These flags were primarily included for exploratory analysis and characterization. They are not used in any way in the work described in the main paper.

Different words that refer to the product:

- *anyone have AE {logs} ?*
  Pm me , i need some {links} will pay 100$ each

- *MySQLi {Dumper} is the best GUI {tool} dedicated to SQL injection attacks on MySQL.*

- *250 .de {rdps}*
  {servers} and xp, more {servers}, some 2008 {servers}, some windows 7, good speeds, mostly admin access
  mostly have usernames and passwords different than whats on the .ru markets

Multiple products:

- *{SQLi} and Admin {Login} for site with 4.8+ million users*

- *need small {hosting} space*
  i need small {hosting} space (3-5gb) + one mysql {base} to host my php stealer - i wont host exe's or
  someother shits, only logs. post your price

- *Buy complete c++ sources of svchost {injection} and AV {bypass}*
  Hi.
  I need you c++ sources - completed, tested and good commented.
  Need work and stable AV {bypass} code in usermode, for example undetectable svchost process injection.