The U.S. FBI Internet Crime Complaint Center was established in 2000 to receive complaints of Internet crime. They produce an annual report, [just released 2017’s edition](https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-the-ic3-2017-internet-crime-report-and-calls-for-increased-public-awareness), and I need the data from it. Since I have to wrangle it out, I thought some folks might like to play long at home, especially since it turns out I had to use both tabulizer and pdftools to accomplish my goal.

Concepts presented:

* PDF scraping (with both tabulizer and pdftools)
* asciiruler
* general string manipulation
* case\_when() vs ifelse() for text cleanup
* reformatting data for ggraph treemaps

Let’s get started! (NOTE: you can click/tap on any image for a larger version)

library(stringi)

library(pdftools)

library(tabulizer)

library(igraph)

library(ggraph) # devtools::install\_github("thomasp85/ggraph")

library(hrbrthemes)

library(tidyverse)

ic3\_file <- "~/Data/2017-ic3-report.pdf" # change "~/Data" for your system

if (!file.exists(ic3\_file)) { # don't waste anyone's bandwidth

download.file("https://pdf.ic3.gov/2017\_IC3Report.pdf", ic3\_file)

}

Let's try pdftools since I like text wrangling

cat(pdftools::pdf\_text(ic3\_file)[[20]])

## 2017 Internet Crime Report 20

## 2017 Crime Types

## By Victim Count

## Crime Type Victims Crime Type Victims

## Non-Payment/Non-Delivery 84,079 Misrepresentation 5,437

## Personal Data Breach 30,904 Corporate Data Breach 3,785

## Phishing/Vishing/Smishing/Pharming 25,344 Investment 3,089

## Overpayment 23,135 Malware/Scareware/Virus 3,089

## No Lead Value 20,241 Lottery/Sweepstakes 3,012

## Identity Theft 17,636 IPR/Copyright and 2,644

## Counterfeit

## Advanced Fee 16,368 Ransomware 1,783

## Harassment/Threats of Violence 16,194 Crimes Against Children 1,300

## Employment 15,784 Denial of Service/TDoS 1,201

## BEC/EAC 15,690 Civil Matter 1,057

## Confidence Fraud/Romance 15,372 Re-shipping 1,025

## Credit Card Fraud 15,220 Charity 436

## Extortion 14,938 Health Care Related 406

## Other 14,023 Gambling 203

## Tech Support 10,949 Terrorism 177

## Real Estate/Rental 9,645 Hacktivist 158

## Government Impersonation 9,149

## Descriptors\*

## Social Media 19,986 \*These descriptors relate to the medium or

## Virtual Currency 4,139 tool used to facilitate the crime, and are used

## by the IC3 for tracking purposes only. They

## are available only after another crime type

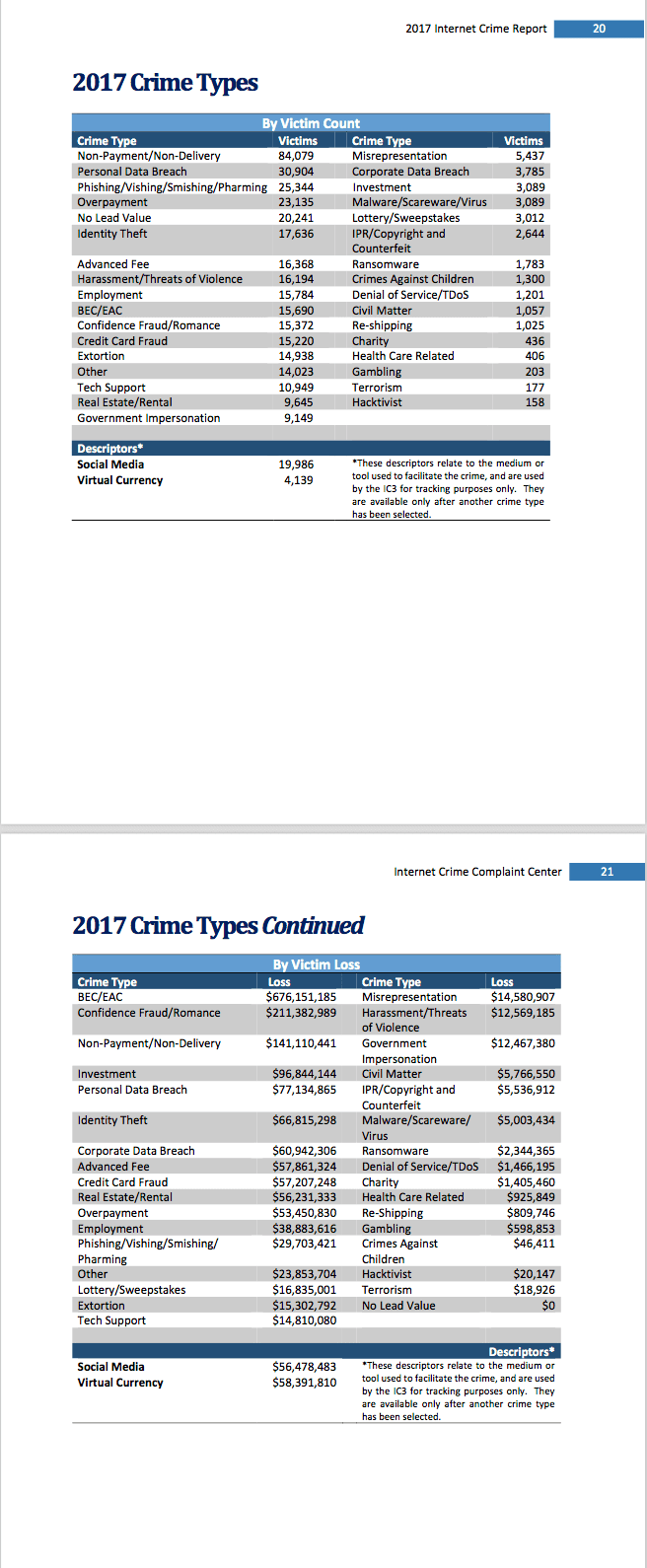
## has been selected.

OK, I don't like text wrangling that much. How about tabulizer?

tabulizer::extract\_tables(ic3\_file, pages = 20)

## list()

Well, that's disappointing. Perhaps if we target the tables on the PDF pages we'll have better luck. You can find them on pages 20 and 21 if you downloaded your own copy. Here are some smaller, static views of them:

[](https://i0.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3-2017-pgs.png?ssl=1)

I can't show the tabulizer pane (well I *could* if I had time to screen capture and make an animated gif) but run this to get the areas:

areas <- tabulizer::locate\_areas(ic3\_file, pages = 20:21)

# this is what ^^ produces for my rectangles:

list(

c(top = 137.11911357341, left = 66.864265927978, bottom = 413.5512465374, right = 519.90581717452),

c(top = 134.92520775623, left = 64.670360110803, bottom = 458.52631578947, right = 529.7783933518)

) -> areas

Now, see if tabulizer can do a better job. We'll start with the first page:

tab <- tabulizer::extract\_tables(ic3\_file, pages = 20, area = areas[1])

tab

## [[1]]

## [,1] [,2]

## [1,] "" "By Victim Cou nt"

## [2,] "Crime Type" "Victims"

## [3,] "Non-Payment/Non-Delivery" "84,079"

## [4,] "Personal Data Breach" "30,904"

## [5,] "Phishing/Vishing/Smishing/Pharming" "25,344"

## [6,] "Overpayment" "23,135"

## [7,] "No Lead Value" "20,241"

## [8,] "Identity Theft" "17,636"

## [9,] "" ""

## [10,] "Advanced Fee" "16,368"

## [11,] "Harassment/Threats of Violence" "16,194"

## [12,] "Employment" "15,784"

## [13,] "BEC/EAC" "15,690"

## [14,] "Confidence Fraud/Romance" "15,372"

## [15,] "Credit Card Fraud" "15,220"

## [16,] "Extortion" "14,938"

## [17,] "Other" "14,023"

## [18,] "Tech Support" "10,949"

## [19,] "Real Estate/Rental" "9,645"

## [20,] "G overnment Impersonation" "9,149"

## [21,] "" ""

## [22,] "Descriptors\*" ""

## [,3] [,4]

## [1,] "" ""

## [2,] "Crime Type" "Victims"

## [3,] "Misrepresentation" "5,437"

## [4,] "Corporate Data Breach" "3,785"

## [5,] "Investment" "3,089"

## [6,] "Malware/Scareware/Virus" "3,089"

## [7,] "Lottery/Sweepstakes" "3,012"

## [8,] "IPR/Copyright and" "2,644"

## [9,] "Counterfeit" ""

## [10,] "Ransomware" "1,783"

## [11,] "Crimes Against Children" "1,300"

## [12,] "Denial of Service/TDoS" "1,201"

## [13,] "Civil Matter" "1,057"

## [14,] "Re-shipping" "1,025"

## [15,] "Charity" "436"

## [16,] "Health Care Related" "406"

## [17,] "Gambling" "203"

## [18,] "Terrorism" "177"

## [19,] "Hacktivist" "158"

## [20,] "" ""

## [21,] "" ""

## [22,] "" ""

Looking good. How does it look data-frame'd?

tab <- as\_data\_frame(tab[[1]])

print(tab, n=50)

## # A tibble: 22 x 4

## V1 V2 V3 V4

## 1 "" By Victim Cou nt "" ""

## 2 Crime Type Victims Crime Type Vict…

## 3 Non-Payment/Non-Delivery 84,079 Misrepresent… 5,437

## 4 Personal Data Breach 30,904 Corporate Da… 3,785

## 5 Phishing/Vishing/Smishing/Pharming 25,344 Investment 3,089

## 6 Overpayment 23,135 Malware/Scar… 3,089

## 7 No Lead Value 20,241 Lottery/Swee… 3,012

## 8 Identity Theft 17,636 IPR/Copyrigh… 2,644

## 9 "" "" Counterfeit ""

## 10 Advanced Fee 16,368 Ransomware 1,783

## 11 Harassment/Threats of Violence 16,194 Crimes Again… 1,300

## 12 Employment 15,784 Denial of Se… 1,201

## 13 BEC/EAC 15,690 Civil Matter 1,057

## 14 Confidence Fraud/Romance 15,372 Re-shipping 1,025

## 15 Credit Card Fraud 15,220 Charity 436

## 16 Extortion 14,938 Health Care … 406

## 17 Other 14,023 Gambling 203

## 18 Tech Support 10,949 Terrorism 177

## 19 Real Estate/Rental 9,645 Hacktivist 158

## 20 G overnment Impersonation 9,149 "" ""

## 21 "" "" "" ""

## 22 Descriptors\* "" "" ""

Still pretty good. Cleaning it up is pretty simple from here. Just filter out some rows, parse some numbers, fix some chopped labels and boom - done:

tab <- filter(tab[3:21,], !V2 == "")

bind\_rows(

select(tab, crime = V1, n\_victims = V2),

select(tab, crime = V3, n\_victims = V4)

) %>%

filter(crime != "") %>%

mutate(n\_victims = readr::parse\_number(n\_victims)) %>%

mutate(crime = case\_when(

stri\_detect\_fixed(crime, "G o") ~ "Government Impersonation",

stri\_detect\_fixed(crime, "IPR/C") ~ "IPR/Copyright and Counterfeit",

TRUE ~ crime

)) %>%

print(n=50) -> ic3\_2017\_crimes\_victim\_ct

## # A tibble: 33 x 2

## crime n\_victims

##

## 1 Non-Payment/Non-Delivery 84079.

## 2 Personal Data Breach 30904.

## 3 Phishing/Vishing/Smishing/Pharming 25344.

## 4 Overpayment 23135.

## 5 No Lead Value 20241.

## 6 Identity Theft 17636.

## 7 Advanced Fee 16368.

## 8 Harassment/Threats of Violence 16194.

## 9 Employment 15784.

## 10 BEC/EAC 15690.

## 11 Confidence Fraud/Romance 15372.

## 12 Credit Card Fraud 15220.

## 13 Extortion 14938.

## 14 Other 14023.

## 15 Tech Support 10949.

## 16 Real Estate/Rental 9645.

## 17 Government Impersonation 9149.

## 18 Misrepresentation 5437.

## 19 Corporate Data Breach 3785.

## 20 Investment 3089.

## 21 Malware/Scareware/Virus 3089.

## 22 Lottery/Sweepstakes 3012.

## 23 IPR/Copyright and Counterfeit 2644.

## 24 Ransomware 1783.

## 25 Crimes Against Children 1300.

## 26 Denial of Service/TDoS 1201.

## 27 Civil Matter 1057.

## 28 Re-shipping 1025.

## 29 Charity 436.

## 30 Health Care Related 406.

## 31 Gambling 203.

## 32 Terrorism 177.

## 33 Hacktivist 158.

Now, on to page 2!

tab <- tabulizer::extract\_tables(ic3\_file, pages = 21, area = areas[2])

tab

## [[1]]

## [,1] [,2]

## [1,] "" "By Victim Lo ss"

## [2,] "Crime Type" "Loss Crime Type"

## [3,] "BEC/EAC" "$676,151,185 Misrepresentation"

## [4,] "Confidence Fraud/Romance" "$211,382,989 Harassment/Threats"

## [5,] "" "of Violence"

## [6,] "Non-Payment/Non-Delivery" "$141,110,441 Government"

## [7,] "" "Impersonation"

## [8,] "Investment" "$96,844,144 Civil Matter"

## [9,] "Personal Data Breach" "$77,134,865 IPR/Copyright and"

## [10,] "" "Counterfeit"

## [11,] "Identity Theft" "$66,815,298 Malware/Scareware/"

## [12,] "" "Virus"

## [13,] "Corporate Data Breach" "$60,942,306 Ransomware"

## [14,] "Advanced Fee" "$57,861,324 Denial of Service/TDoS"

## [15,] "Credit Card Fraud" "$57,207,248 Charity"

## [16,] "Real Estate/Rental" "$56,231,333 Health Care Related"

## [17,] "Overpayment" "$53,450,830 Re-Shipping"

## [18,] "Employment" "$38,883,616 Gambling"

## [19,] "Phishing/Vishing/Smishing/" "$29,703,421 Crimes Against"

## [20,] "Pharming" "Children"

## [21,] "Other" "$23,853,704 Hacktivist"

## [22,] "Lottery/Sweepstakes" "$16,835,001 Terrorism"

## [23,] "Extortion" "$15,302,792 N o Lead Value"

## [24,] "Tech Support" "$14,810,080"

## [25,] "" ""

## [26,] "" ""

## [,3]

## [1,] ""

## [2,] "Loss"

## [3,] "$14,580,907"

## [4,] "$12,569,185"

## [5,] ""

## [6,] "$12,467,380"

## [7,] ""

## [8,] "$5,766,550"

## [9,] "$5,536,912"

## [10,] ""

## [11,] "$5,003,434"

## [12,] ""

## [13,] "$2,344,365"

## [14,] "$1,466,195"

## [15,] "$1,405,460"

## [16,] "$925,849"

## [17,] "$809,746"

## [18,] "$598,853"

## [19,] "$46,411"

## [20,] ""

## [21,] "$20,147"

## [22,] "$18,926"

## [23,] "$0"

## [24,] ""

## [25,] ""

## [26,] "Descriptors\*"

:facepalm: That's disappointing. Way too much scrambled content. So, back to pdftools!

cat(pg21 <- pdftools::pdf\_text(ic3\_file)[[21]])

## Internet Crime Complaint Center 21

## 2017 Crime Types Continued

## By Victim Loss

## Crime Type Loss Crime Type Loss

## BEC/EAC $676,151,185 Misrepresentation $14,580,907

## Confidence Fraud/Romance $211,382,989 Harassment/Threats $12,569,185

## of Violence

## Non-Payment/Non-Delivery $141,110,441 Government $12,467,380

## Impersonation

## Investment $96,844,144 Civil Matter $5,766,550

## Personal Data Breach $77,134,865 IPR/Copyright and $5,536,912

## Counterfeit

## Identity Theft $66,815,298 Malware/Scareware/ $5,003,434

## Virus

## Corporate Data Breach $60,942,306 Ransomware $2,344,365

## Advanced Fee $57,861,324 Denial of Service/TDoS $1,466,195

## Credit Card Fraud $57,207,248 Charity $1,405,460

## Real Estate/Rental $56,231,333 Health Care Related $925,849

## Overpayment $53,450,830 Re-Shipping $809,746

## Employment $38,883,616 Gambling $598,853

## Phishing/Vishing/Smishing/ $29,703,421 Crimes Against $46,411

## Pharming Children

## Other $23,853,704 Hacktivist $20,147

## Lottery/Sweepstakes $16,835,001 Terrorism $18,926

## Extortion $15,302,792 No Lead Value $0

## Tech Support $14,810,080

## Descriptors\*

## Social Media $56,478,483 \*These descriptors relate to the medium or

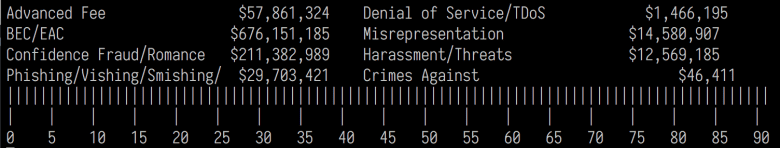
## Virtual Currency $58,391,810 tool used to facilitate the crime, and are used

## by the IC3 for tracking purposes only. They

## are available only after another crime type

## has been selected.

This is (truthfully) not too bad. Just make columns from substring ranges and do some cleanup. The asciiruler package can definitely help here since it makes it much easier to see start/stop points (I used a new editor pane and copied some lines into it):

[](https://i1.wp.com/rud.is/b/wp-content/uploads/2018/05/asciiruler.png?ssl=1)

stri\_split\_lines(pg21)[[1]] %>%

.[-(1:4)] %>% # remove header & bits above header

.[-(26:30)] %>% # remove trailing bits

map\_df(~{

list(

crime = stri\_trim\_both(c(stri\_sub(.x, 1, 25), stri\_sub(.x, 43, 73))),

cost = stri\_trim\_both(c(stri\_sub(.x, 27, 39), stri\_sub(.x, 74))) # no length/to in the last one so it goes until eol

)

}) %>%

filter(!(crime == "" | cost == "")) %>% # get rid of blank rows

mutate(cost = suppressWarnings(readr::parse\_number(cost))) %>% # we can use NAs generated to remove non-data rows

filter(!is.na(cost)) %>%

mutate(crime = case\_when(

stri\_detect\_fixed(crime, "Phish") ~ "Phishing/Vishing/Smishing/Pharming",

stri\_detect\_fixed(crime, "Malware") ~ "Malware/Scareware/Virus",

stri\_detect\_fixed(crime, "IPR") ~ "IPR/Copyright and Counterfeit",

stri\_detect\_fixed(crime, "Harassment") ~ "Harassment/Threats of Violence",

TRUE ~ crime

)) %>%

print(n=50) -> ic3\_2017\_crimes\_cost

## # A tibble: 35 x 2

## crime cost

## 1 BEC/EAC 676151185.

## 2 Misrepresentation 14580907.

## 3 Confidence Fraud/Romance 211382989.

## 4 Harassment/Threats of Violence 12569185.

## 5 Non-Payment/Non-Delivery 141110441.

## 6 Government 12467380.

## 7 Investment 96844144.

## 8 Civil Matter 5766550.

## 9 Personal Data Breach 77134865.

## 10 IPR/Copyright and Counterfeit 5536912.

## 11 Identity Theft 66815298.

## 12 Malware/Scareware/Virus 5003434.

## 13 Corporate Data Breach 60942306.

## 14 Ransomware 2344365.

## 15 Advanced Fee 57861324.

## 16 Denial of Service/TDoS 1466195.

## 17 Credit Card Fraud 57207248.

## 18 Charity 1405460.

## 19 Real Estate/Rental 56231333.

## 20 Health Care Related 925849.

## 21 Overpayment 53450830.

## 22 Re-Shipping 809746.

## 23 Employment 38883616.

## 24 Gambling 598853.

## 25 Phishing/Vishing/Smishing/Pharming 29703421.

## 26 Crimes Against 46411.

## 27 Other 23853704.

## 28 Hacktivist 20147.

## 29 Lottery/Sweepstakes 16835001.

## 30 Terrorism 18926.

## 31 Extortion 15302792.

## 32 No Lead Value 0.

## 33 Tech Support 14810080.

## 34 Social Media 56478483.

## 35 Virtual Currency 58391810.

Now that we have real data, we can take a look at the IC3 crimes by loss and victims.

We'll use treemaps first then take a quick look at the relationship between counts and losses.

Just need to do some data wrangling for ggraph, starting with victims:

ic3\_2017\_crimes\_victim\_ct %>%

mutate(crime = case\_when(

crime == "Government Impersonation" ~ "Government\nImpersonation",

crime == "Corporate Data Breach" ~ "Corporate\nData\nBreach",

TRUE ~ crime

)) %>%

mutate(crime = stri\_replace\_all\_fixed(crime, "/", "/\n")) %>%

mutate(grp = "ROOT") %>%

add\_row(grp = "ROOT", crime="ROOT", n\_victims=0) %>%

select(grp, crime, n\_victims) %>%

arrange(desc(n\_victims)) -> gdf

select(gdf, -grp) %>%

mutate(lab = sprintf("%s\n(%s)", crime, scales::comma(n\_victims))) %>%

mutate(lab = ifelse(n\_victims > 1300, lab, "")) %>% # don't show a label when blocks are small

mutate(lab\_col = ifelse(n\_victims > 40000, "#2b2b2b", "#cccccc")) -> vdf # change up colors when blocks are lighter

g <- graph\_from\_data\_frame(gdf, vertices=vdf)

ggraph(g, "treemap", weight=n\_victims) +

geom\_node\_tile(aes(fill=n\_victims), size=0.25) +

geom\_text(

aes(x, y, label=lab, size=n\_victims, color = I(lab\_col)),

family=font\_ps, lineheight=0.875

) +

scale\_x\_reverse(expand=c(0,0)) +

scale\_y\_continuous(expand=c(0,0)) +

scale\_size\_continuous(trans = "sqrt", range = c(0.5, 8)) +

labs(title = "FBI 2017 Internet Crime Report — Crimes By Victim Count") +

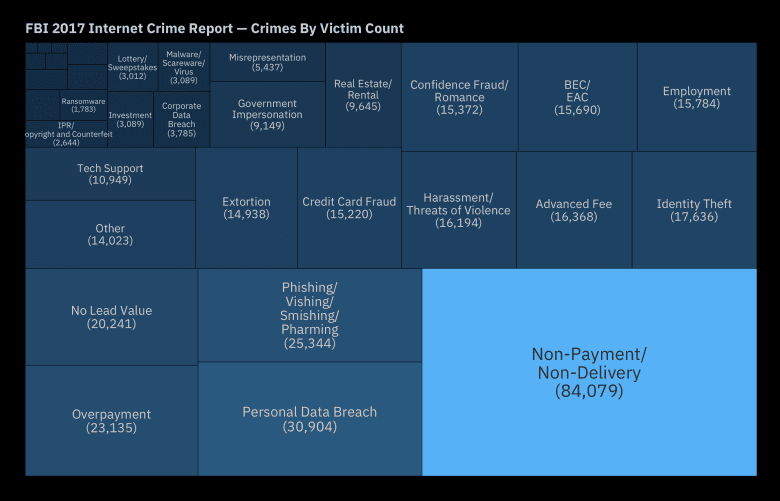
ggraph::theme\_graph(base\_family = font\_ps) +

theme(plot.title = element\_text(color="#cccccc", family = "IBMPlexSans-Bold")) +

theme(panel.background = element\_rect(fill="black")) +

theme(plot.background = element\_rect(fill="black")) +

theme(legend.position="none")

[](https://i0.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3_victim_treemap-1.png?ssl=1)

# Now, do the same for losses:

ic3\_2017\_crimes\_cost %>%

mutate(crime = case\_when(

crime == "Phishing/Vishing/Smishing/Pharming" ~ "Phishing/Vishing/\nSmishing/Pharming",

crime == "Harassment/Threats of Violence" ~ "Harassment/\nThreats of Violence",

crime == "Lottery/Sweepstakes" ~ "Lottery\nSweepstakes",

TRUE ~ crime

)) %>%

filter(cost > 0) %>%

mutate(cost = cost / 1000000) %>%

mutate(grp = "ROOT") %>%

add\_row(grp = "ROOT", crime="ROOT", cost=0) %>%

select(grp, crime, cost) %>%

arrange(desc(cost)) -> gdf\_cost

select(gdf\_cost, -grp) %>%

mutate(lab = sprintf("%s\n($%s M)", crime, prettyNum(cost, digits=2))) %>%

mutate(lab = ifelse(cost > 5.6, lab, "")) %>%

mutate(lab\_col = ifelse(cost > 600, "#2b2b2b", "#cccccc")) -> vdf\_cost

g\_cost <- graph\_from\_data\_frame(gdf\_cost, vertices=vdf\_cost)

ggraph(g\_cost, "treemap", weight=cost) +

geom\_node\_tile(aes(fill=cost), size=0.25) +

geom\_text(

aes(x, y, label=lab, size=cost, color=I(lab\_col)),

family=font\_ps, lineheight=0.875

) +

scale\_x\_reverse(expand=c(0,0)) +

scale\_y\_continuous(expand=c(0,0)) +

scale\_size\_continuous(trans = "sqrt", range = c(0.5, 8)) +

labs(title = "FBI 2017 Internet Crime Report — Crime Loss By Category") +

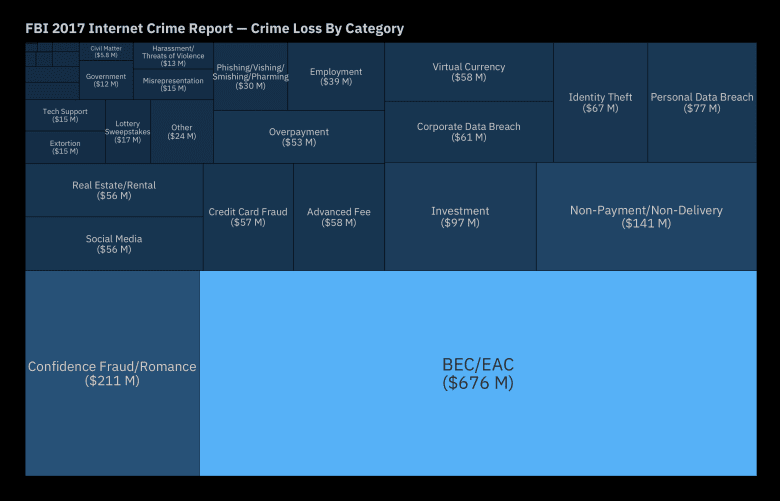
ggraph::theme\_graph(base\_family = font\_ps) +

theme(plot.title = element\_text(color="#cccccc", family = "IBMPlexSans-Bold")) +

theme(panel.background = element\_rect(fill="black")) +

theme(plot.background = element\_rect(fill="black")) +

theme(legend.position="none")

[](https://i1.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3_crimes_loss-1.png?ssl=1)

Let's plot victim counts vs losses to see what stands out:

left\_join(ic3\_2017\_crimes\_victim\_ct, ic3\_2017\_crimes\_cost) %>%

filter(!is.na(cost)) %>%

ggplot(aes(n\_victims, cost)) +

geom\_point() +

ggrepel::geom\_label\_repel(aes(label = crime), family=font\_ps, size=3) +

scale\_x\_comma() +

scale\_y\_continuous(labels=scales::dollar) +

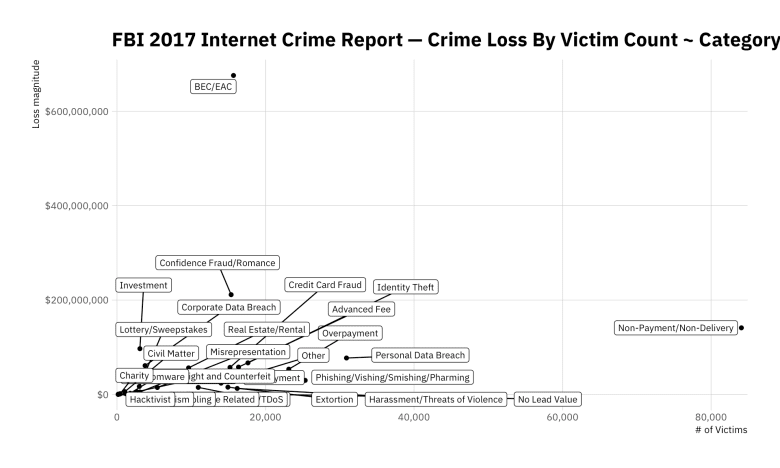
labs(

x = "# of Victims", y = "Loss magnitude",

title = "FBI 2017 Internet Crime Report — Crime Loss By Victim Count ~ Category"

) +

theme\_ipsum\_ps(grid="XY")

[](https://i1.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3_victim_loss_1-1.png?ssl=1)

BEC == "[Business email compromise](https://www.fbi.gov/news/stories/business-e-mail-compromise-on-the-rise)" and it's definitely a major problem, but those two count/loss outliers are not helping us see the rest of the data. Let's zoom in:

left\_join(ic3\_2017\_crimes\_victim\_ct, ic3\_2017\_crimes\_cost) %>%

filter(!is.na(cost)) %>%

filter(cost < 300000000) %>%

filter(n\_victims < 40000) %>%

ggplot(aes(n\_victims, cost)) +

geom\_point() +

ggrepel::geom\_label\_repel(aes(label = crime), family=font\_ps, size=3) +

scale\_x\_comma() +

scale\_y\_continuous(labels=scales::dollar) +

labs(

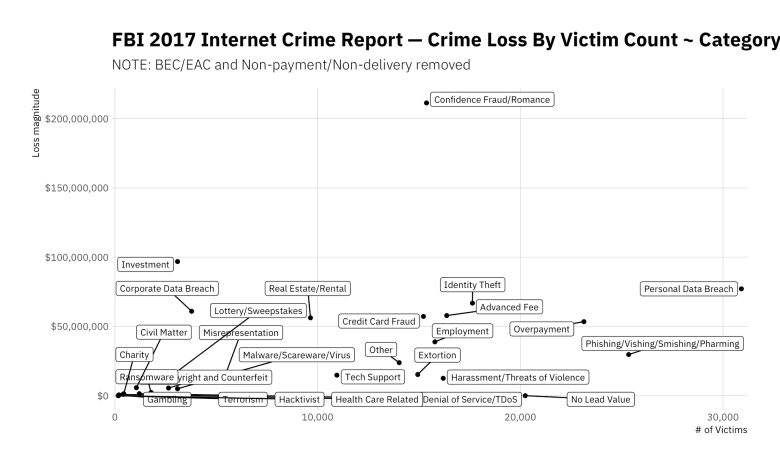
x = "# of Victims", y = "Loss magnitude",

title = "FBI 2017 Internet Crime Report — Crime Loss By Victim Count ~ Category",

subtitle = "NOTE: BEC/EAC and Non-payment/Non-delivery removed"

) +

theme\_ipsum\_ps(grid="XY")

[](https://i2.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3_victim_loss_2-1.png?ssl=1)

Better, but let's go zoom in a bit more:

left\_join(ic3\_2017\_crimes\_victim\_ct, ic3\_2017\_crimes\_cost) %>%

filter(!is.na(cost)) %>%

filter(cost < 50000000) %>%

filter(n\_victims < 10000) %>%

ggplot(aes(n\_victims, cost)) +

geom\_point() +

ggrepel::geom\_label\_repel(aes(label = crime), family=font\_ps, size=3) +

scale\_x\_comma() +

scale\_y\_continuous(labels=scales::dollar) +

labs(

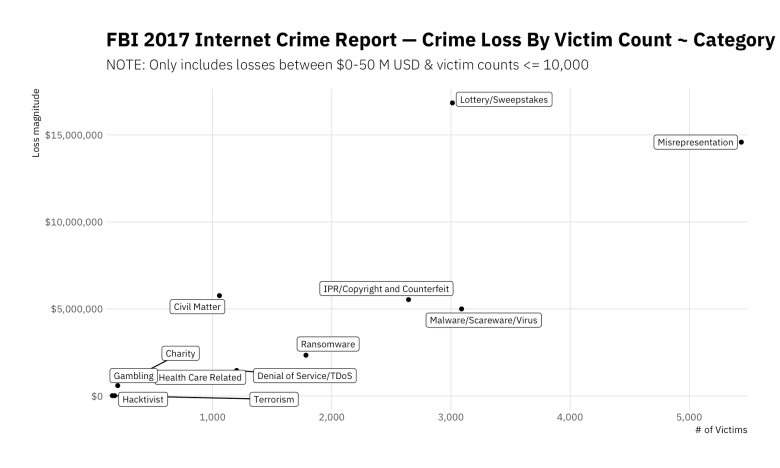
x = "# of Victims", y = "Loss magnitude",

title = "FBI 2017 Internet Crime Report — Crime Loss By Victim Count ~ Category",

subtitle = "NOTE: Only includes losses between $0-50 M USD & victim counts <= 10,000 "

) +

theme\_ipsum\_ps(grid="XY")

[](https://i1.wp.com/rud.is/b/wp-content/uploads/2018/05/ic3_victim_loss_3-1.png?ssl=1)

Looks like the ransomware folks have quite a bit of catching up to do (at least when it comes to crimes reported to the IC3).