Implicit Generative Modeling for Password Estimation

joint work with B. Hitaj, P. Gasti and G. Ateniese

Fernando Perez-Cruz

Swiss Data Science Center (ETH Zurich and EPFL)

Bern Winter School on Machine Learning

SDSC

Introduction

- Passwords are the most popular authentication method (default).
- Password database leaks have shown that users tend to choose easy-to-guess passwords.
- John The Reaper or HashCat:
  - Dictionaries.
  - Previous password leaks.
  - Heuristics for password transformations.
  - Combinations of multiple words (e.g., iloveyou123456).
  - Mixed letter case (e.g., iLoVeyOu).
  - Leet speak (e.g., il0v3you).
- Context Free Grammars.
- Neural Network – FLA.

SDSC

Google Recommendations

Step 1: Create a strong password

A strong password helps you:

- Keep your personal info safe
- Protect your emails, files, and other content
- Prevent someone else from getting in to your account

Meet password requirements

Create your password using 8 characters or more. It can be any combination of letters, numbers, and symbols.

You can't use a password that:

- Is particularly weak. Example: "password123"
- You've used before on your account

Follow tips for a good password

A strong password can be memorable to you but nearly impossible for someone else to guess. Learn what makes a good password, then follow these tips to create your own.

Make your password unique

Make your password longer & more memorable

Use letters, numbers & symbols

SDSC

Google Recommendations

Use letters, numbers & symbols

Passwords with different types of symbols might be more difficult for people to guess, but also might be harder for you to remember. You can set up recovery info to avoid getting locked out if you forget your password.

Combine different types of characters

Use a mix of alphanumeric characters (letters and numbers) and symbols:

- Uppercase (capital) letters. Examples: A, E, R
- Lowercase (small) letters. Examples: a, e, r
- Numbers. Examples: 2, 6, 7
- Symbols and special characters. Examples: ! @ & *

Recommendations & examples

Replace letters with numbers & symbols: Choose a word or phrase and use numbers and symbols instead of some letters. Examples:

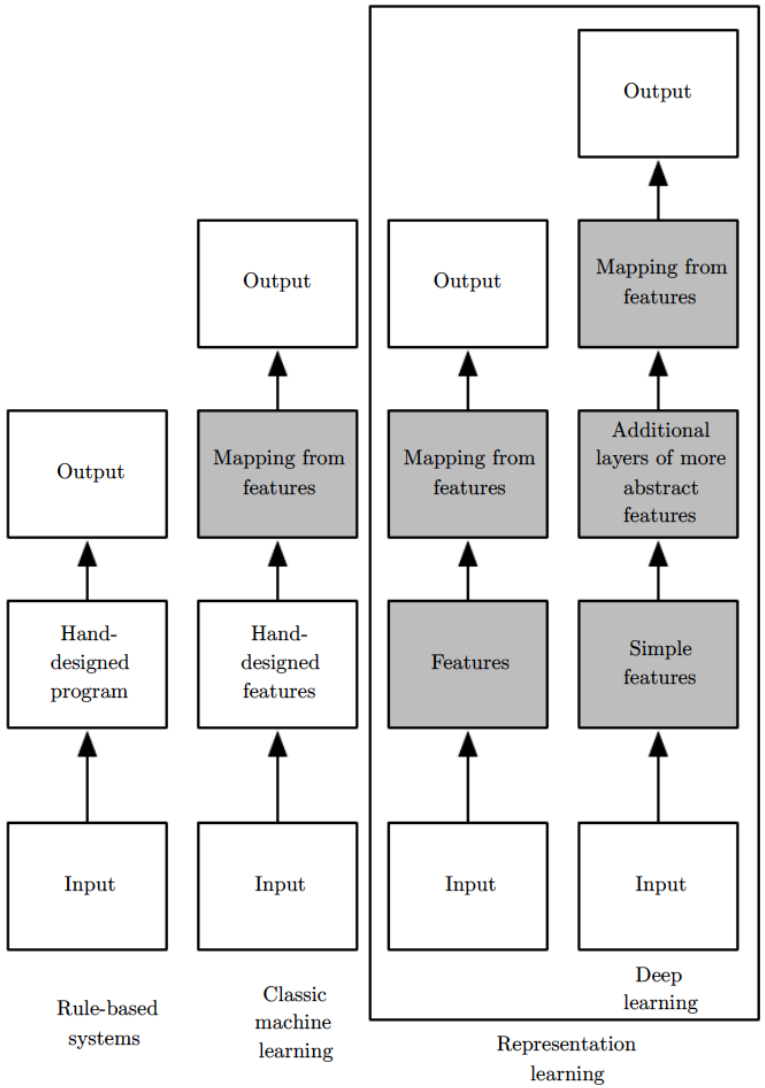
- "Spooky Halloween" becomes "sPo0kyH@ll0w3En"
- "Later gator" becomes "L8rg@+0R"

Abbreviate a sentence: Come up with a sentence and use the first letter of each word. Example:

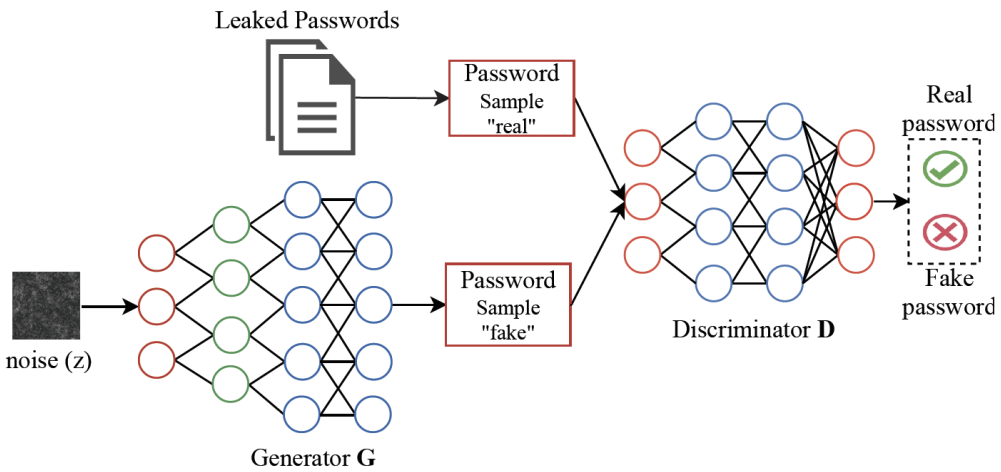
- "Uncle Peter always ate chocolate-covered everything" becomes "uP@8cCe!"

SDSC

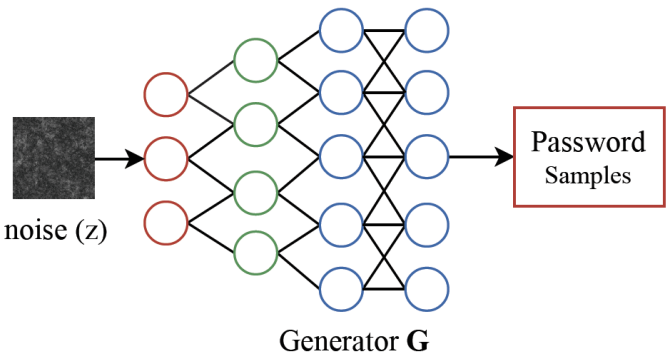
Remember this plot from the morning?



Password learning and generation



Improved Wasserstein Gan



SDSC

General Results

Approach	(1) Unique Passwords	(2) Matches
JTR Spyderlab	10^9	461,395 (23.32%)
Markov Model 3-gram	$4.9 \cdot 10^8$	532,961 (26.93%)
HashCat gen2	10^9	597,899 (30.22%)
HashCat Best64	$3.6 \cdot 10^8$	630,068 (31.84%)
PCFG	10^{10}	650,695 (32.89%)
FLA 10^{-10}	$7.4 \cdot 10^8$	652,585 (32.99%)
PassGAN	$2.1 \cdot 10^9$	515,079 (26.04%)
PassGAN	$3.6 \cdot 10^9$	584,466 (29.54%)
PassGAN	$4.9 \cdot 10^9$	625,245 (31.60%)
PassGAN	$6.0 \cdot 10^9$	653,978 (33.06%)
PassGAN	$7.1 \cdot 10^9$	676,439 (34.19%)

Unique Passwords	(1) PassGAN	(2) FLA	(3) PassGAN ∪ FLA	(4) PassGAN, and not from FLA	(5) FLA, and not from PassGAN
10^4	14	2	16	14	2
10^5	95	40	133	93	38
10^6	881	1,183	2,016	833	1,135
10^7	7,633	16,330	22,203	5,873	14,570
10^8	44,490	117,262	137,415	20,153	92,925
10^9	155,369				
$7 \cdot 10^9$	320,365				

High probability passwords

Password	Occurrence in Training Data	Frequency in Training Data	GAN Estimated Frequency
123456	232,844	0.98%	100,971,288 (1.0%)
12345	63,135	0.27%	21,614,548 (0.22%)
123456789	61,531	0.26%	22,208,040 (0.22%)
password	47,507	0.20%	85,889 (8.6e-4%)
iloveyou	40,037	0.17%	10,056,700 (0.10%)
princess	26,669	0.11%	190,796 (0.0019%)
1234567	17,399	0.073%	7,545,708 (0.075%)
rockyou	16,765	0.071%	55,515 (5.5e-4%)
12345678	16,536	0.070%	5,070,673 (0.051%)
abc123	13,243	0.056%	6,545 (6.5e-5%)
nicole	12,992	0.055%	206,277 (0.0021%)
daniel	12,337	0.052%	3,304,567 (0.033%)
babygirl	12,130	0.051%	13,076 (1.3e-4%)
monkey	11,726	0.050%	116,602 (0.0012%)
lovely	11,533	0.049%	1,026,362 (0.010%)
jessica	11,262	0.048%	220,849 (0.0022%)
654321	11,181	0.047%	19,912 (1.9e-4%)
michael	11,174	0.047%	517 (5.2e-6%)
ashley	10,741	0.045%	116,858 (0.0012%)
qwerty	10,730	0.045%	135,124 (0.0013%)
iloveu	10,587	0.045%	4,839,368 (0.048%)
111111	10,529	0.044%	101,903 (0.0010%)
000000	10,412	0.044%	108,300 (0.0011%)
michelle	10,210	0.043%	739,220 (0.0073%)
tigger	9,381	0.040%	658,360 (0.0066%)
sunshine	9,252	0.039%	3,628 (3.6e-5%)
chocolate	9,012	0.038%	12 (1.2e-7%)
password1	8,916	0.038%	6,427 (6.4e-5%)
soccer	8,752	0.037%	25 (2.5e-7%)
anthony	8,752	0.036%	not generated

SDSC

High probability passwords

Password	Rank in Training Set	Frequency in Training Set	Probability assigned by FLA
123456	1	0.9833%	2.81E-3
12345	2	0.2666%	1.06E-3
123457	3,224	0.0016%	2.87E-4
1234566	5,769	0.0010%	1.85E-4
1234565	9,692	0.0006%	1.11E-4
1234567	7	0.0735%	1.00E-4
12345669	848,078	0.0000%	9.84E-5
123458	7,359	0.0008%	9.54E-5
12345679	7,818	0.0007%	9.07E-5
123459	8,155	0.0007%	7.33E-5
lover	457	0.0079%	6.73E-5
love	384	0.0089%	6.09E-5
223456	69,163	0.0001%	5.14E-5
22345	118,098	0.0001%	4.61E-5
1234564	293,340	0.0000%	3.81E-5

FLA

Password	Rank in Training Set	Frequency in Training Set	Probability in PassGAN's Output
123456	1	0.9833%	1.01E-2
123456789	3	0.25985%	2.2E-3
12345	2	0.26662%	2.16E-3
iloveyou	5	0.16908%	1.01E-3
1234567	7	0.07348%	7.6E-4
angel	33	0.03558%	6.4E-4
12345678	9	0.06983%	5.1E-4
iloveu	21	0.04471%	4.9E-4
angela	109	0.01921%	3.4E-4
daniel	12	0.0521%	3.3E-4
sweety	90	0.02171%	2.6E-4
angels	57	0.02787%	2.5E-4
maria	210	0.01342%	1.6E-4
loveyou	52	0.0287%	1.5E-4
andrew	55	0.02815%	1.3E-4

GAN

SDSC

Conclusion about PassGANs

- PassGAN is comparable to current Passwords guessing procedures.
- PassGANs provides better density estimate (not good anyway).
- New structures are needed to improve in the mid-range estimates.
- Passwords is a good benchmark for text generation GANs:
  - It is a relevant problem.
  - It is easier than natural language.
  - It has an *on-the-job* performance measure: number of guess passwords.
  - It can help with density estimation evaluation, as the choices are discrete and repeatable.

SDSC